Entropy **2015**, *17*, 4644-4653; doi:10.3390/e17074644

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Quantifying Redundant Information in Predicting a Target Random Variable

Virgil Griffith [1,*] and Tracey Ho [2]

[1] School of Computing, National University of Singapore, Singapore 119077, Singapore
[2] Computer Science and Electrical Engineering, Caltech, Pasadena, CA 91125, USA;
E-Mail: tho@caltech.edu

* Author to whom correspondence should be addressed; E-Mail: i@virgil.gr.

Academic Editor: Rick Quax

Received: 18 March 2015 / Accepted: 26 June 2015 / Published: 2 July 2015

Abstract: We consider the problem of defining a measure of redundant information that quantifies how much common information two or more random variables specify about a target random variable. We discussed desired properties of such a measure, and propose new measures with some desirable properties.

Keywords: synergy; information theory; complex systems; irreducibility; synergistic information; intersection-information

1. Introduction

Many molecular and neurological systems involve multiple interacting factors affecting an outcome synergistically and/or redundantly. Attempts to shed light on issues such as population coding in neurons, or genetic contribution to a phenotype (e.g., eye-color), have motivated various proposals to leverage principled information-theoretic measures for quantifying informational synergy and redundancy, e.g., [1–5]. In these settings, we are concerned with the statistics of how two (or more) random variables X_1, X_2, called predictors, jointly or separately specify/predict another random variable Y, called a target random variable. This focus on a target random variable is in contrast to Shannon's mutual information which quantifies statistical dependence between two random variables, and various notions of common information, e.g., [6–8].

The concepts of synergy and redundancy are based on several intuitive notions, e.g., positive informational synergy indicates that X_1 and X_2 act cooperatively or antagonistically to influence Y; positive redundancy indicates there is an aspect of Y that X_1 and X_2 can each separately predict. However, it has been challenging [9–12] to come up with precise information-theoretic definitions of synergy and redundancy that are consistent with all intuitively desired properties.

2. Background: Partial Information Decomposition

Partial Information Decomposition (PID) [13] defines the concepts of synergistic, redundant and unique information in terms of *intersection information*, $\mathrm{I}_\cap(\{X_1, \ldots, X_n\} : Y)$, which quantifies the common information that each of the n predictors $X_1, \ldots, X_n$ conveys about a target random variable Y. An antichain lattice [14] of redundant, unique, and synergistic partial informations is built from the intersection information.

Partial information diagrams (PI-diagrams) extend Venn diagrams to represent synergy. A PI-diagram is composed of nonnegative *partial information regions* (PI-regions). Unlike the standard Venn entropy diagram in which the sum of all regions is the joint entropy $\mathrm{H}(X_{1\ldots n}, Y)$, in PI-diagrams the sum of all regions (i.e. the space of the PI-diagram) is the mutual information $\mathrm{I}(X_{1\ldots n} : Y)$. PI-diagrams show how the mutual information $\mathrm{I}(X_{1\ldots n} : Y)$ is distributed across subsets of the predictors. For example, in the PI-diagram for $n = 2$ (Figure 1): $\{1\}$ denotes the unique information about Y that only X_1 carries (likewise $\{2\}$ denotes the information only X_2 carries); $\{1, 2\}$ denotes the redundant information about Y that X_1 as well as X_2 carries, while $\{12\}$ denotes the information about Y that is specified only by X_1 and X_2 synergistically or jointly.

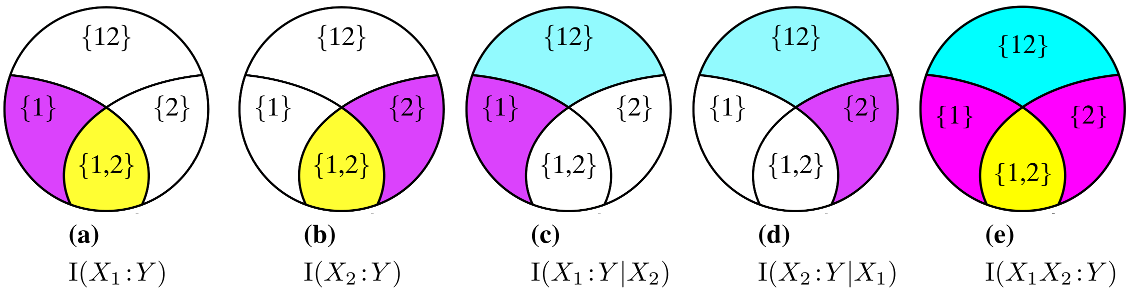


Figure 1. PI-diagrams for $n = 2$ predictors, showing the amount of redundant (yellow/bottom), unique (magenta/left and right) and synergistic (cyan/top) information with respect to the target Y.

Each PI-region is either redundant, unique, or synergistic, but any combination of positive PI-regions may be possible. Per [13], for two predictors, the four partial informations are defined as follows: the redundant information as $\mathrm{I}_\cap(\{X_1, X_2\} : Y)$, the unique informations as

$$\begin{aligned} \mathrm{I}_\partial(\{X_1\} : Y) &= \mathrm{I}(X_1 : Y) - \mathrm{I}_\cap(\{X_1, X_2\} : Y) \\ \mathrm{I}_\partial(\{X_2\} : Y) &= \mathrm{I}(X_2 : Y) - \mathrm{I}_\cap(\{X_1, X_2\} : Y)\,, \end{aligned} \tag{1}$$

and the synergistic information as

$$\begin{aligned} \mathrm{I}_\partial(\{X_1, X_2\} : Y) &= \mathrm{I}(X_1, X_2 : Y) - \mathrm{I}_\partial(\{X_1\} : Y) - \mathrm{I}_\partial(\{X_2\} : Y) - \mathrm{I}_\partial(\{X_1, X_2\} : Y) \\ &= \mathrm{I}(X_1, X_2 : Y) - \mathrm{I}(X_1 : Y) - \mathrm{I}(X_2 : Y) + \mathrm{I}_\cap(\{X_1, X_2\} : Y)\,. \end{aligned} \tag{2}$$

3. Desired $\mathrm{I}_\cap$ properties and canonical examples

There are a number of intuitive properties, proposed in [5,9–13], that are considered desirable for the intersection information measure $\mathrm{I}_\cap$ to satisfy:

($\mathbf{S_0}$) Weak Symmetry: $\mathrm{I}_\cap(\{X_1,\ldots,X_n\}\!:\!Y)$ is invariant under reordering of $X_1,\ldots,X_n$.

($\mathbf{M_0}$) Weak Monotonicity: $\mathrm{I}_\cap(\{X_1,\ldots,X_n,Z\}\!:\!Y) \leq \mathrm{I}_\cap(\{X_1,\ldots,X_n\}\!:\!Y)$ with equality if there exists $X_i \in \{X_1,\ldots,X_n\}$ such that $\mathrm{H}(Z,X_i)=\mathrm{H}(Z)$.

Weak Monotonicity is a natural generalization of the monotonicity property from [13]. Weak monotonicity is inspired by the property of mutual information that if $\mathrm{H}(X|Z) = 0$, then $\mathrm{I}(X\!:\!Y) \leq \mathrm{I}(Z\!:\!Y)$.

(**SR**) Self-Redundancy: $\mathrm{I}_\cap(\{X_1\}\!:\!Y) = \mathrm{I}(X_1\!:\!Y)$. The intersection information a single predictor X_1 conveys about the target Y is equal to the mutual information between the X_1 and the target Y.

($\mathbf{M_1}$) Strong Monotonicity: $\mathrm{I}_\cap(\{X_1,\ldots,X_n,Z\}\!:\!Y) \leq \mathrm{I}_\cap(\{X_1,\ldots,X_n\}\!:\!Y)$ with equality if there exists $X_i \in \{X_1,\ldots,X_n\}$ such that $\mathrm{I}(Z,X_i\!:\!Y) = \mathrm{I}(Z\!:\!Y)$.

Strong Monotonicity captures more precisely what is meant by "redundant information", it says explicitly that it *information about* Y that is redundant, not just any redundancy among the predictors (weak monotonicity).

(**LP**) Local Positivity: For all n, the derived "partial informations" defined in [13] are nonnegative. This is equivalent to requiring that $\mathrm{I}_\cap$ satisfy *total monotonicity*, a stronger form of supermodularity. For $n=2$ this can be concretized as, $\mathrm{I}_\cap(\{X_1,X_2\}\!:\!Y) \geq \mathrm{I}(X_1\!:\!X_2) - \mathrm{I}(X_1\!:\!X_2|Y)$.

(**TM**) Target Monotonicity: If $\mathrm{H}(Y|Z)=0$, then $\mathrm{I}_\cap(\{X_1,\ldots,X_n\}\!:\!Y) \leq \mathrm{I}_\cap(\{X_1,\ldots,X_n\}\!:\!Z)$.

There are also a number of canonical examples for which one or more of the partial informations have intuitive values, which are considered desirable for the intersection information measure $\mathrm{I}_\cap$ to attain.

Example Unq, shown in Figure 2, is a canonical case of unique information, in which each predictor carries independent information about the target. Y has four equiprobable states: `ab`, `aB`, `Ab`, and `AB`. X_1 uniquely specifies bit `a`/`A`, and X_2 uniquely specifies bit `b`/`B`. Note that the states are named so as to highlight the two bits of unique information; it is equivalent to choose any four unique names for the four states.

Example RdnXor, shown in Figure 3, is a canonical example of redundancy and synergy coexisting. The `r`/`R` bit is redundant, while the 0/1 bit of Y is synergistically specified as the XOR of the corresponding bits in X_1 and X_2.

Example And, shown in Figure 4, is an example where the relationship between X_1, X_2 and Y is nonlinear, making the desired partial information values less intuitively obvious. Nevertheless, it is desired that the partial information values should be nonnegative.

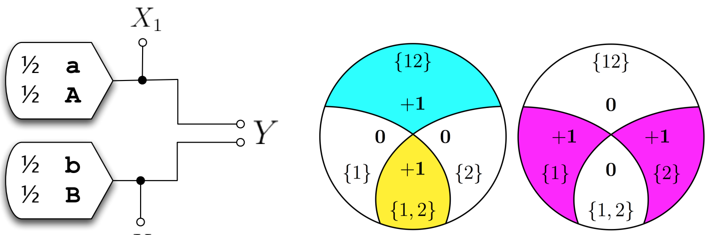
Example ImperfectRdn, shown in Figure 5, is an example of "imperfect" or "lossy" correlation between the predictors, where it is intuitively desirable that the derived redundancy should be positive. Given (**LP**), we can determine the desired decomposition analytically. First, $\mathrm{I}(X_1,X_2\!:\!Y) =$

$\mathrm{I}(X_1 \colon Y) = 1$ bit; therefore, $\mathrm{I}(X_2 \colon Y | X_1) = \mathrm{I}(X_1, X_2 \colon Y) - \mathrm{I}(X_1 \colon Y) = 0$ bits. This determines two of the partial informations—the synergistic information $\mathrm{I}_\partial(\{X_1, X_2\} Y)$ and the unique information $\mathrm{I}_\partial(\{X_2\} : Y)$ are both zero. Then, the redundant information $\mathrm{I}_\partial(\{X_1, X_2\} : Y) = \mathrm{I}(X_2 \colon Y) - \mathrm{I}_\partial(\{X_2\} : Y) = \mathrm{I}(X_2 \colon Y) = 0.99$ bits. Having determined three of the partial informations, we compute the final unique information $\mathrm{I}_\partial(\{X_1\} : Y) = \mathrm{I}(X_1 \colon Y) - 0.99 = 0.01$ bits.

X_1	X_2	Y	
a	b	ab	1/4
a	B	aB	1/4
A	b	Ab	1/4
A	B	AB	1/4

(a) $\Pr(x_1, x_2, y)$

$$\mathrm{I}(X_1, X_2 \colon Y) = 2$$
$$\mathrm{I}(X_1 \colon Y) = 1$$
$$\mathrm{I}(X_2 \colon Y) = 1$$



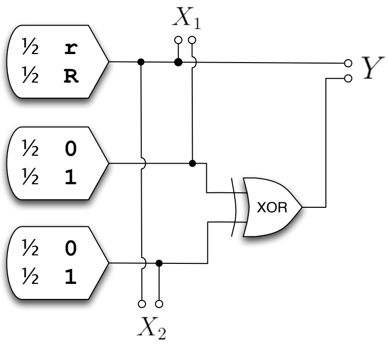
(b) circuit diagram **(c)** $\mathrm{I}_{\min}$ **(d)** WMS/$\mathrm{I}_\wedge$/I_α

Figure 2. Example UNQ. X_1 and X_2 each uniquely carry one bit of information about Y. $\mathrm{I}(X_1 X_2 \colon Y) = \mathrm{H}(Y) = 2$ bits.

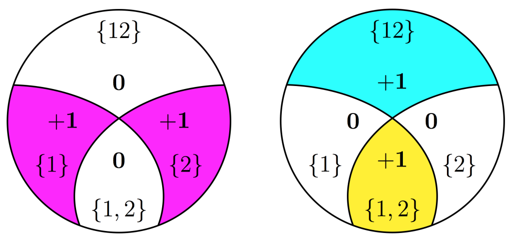
X_1	X_2	Y	
r0	r0	r0	1/8
r0	r1	r1	1/8
r1	r0	r1	1/8
r1	r1	r0	1/8
R0	R0	R0	1/8
R0	R1	R1	1/8
R1	R0	R1	1/8
R1	R1	R0	1/8

(a) $\Pr(x_1, x_2, y)$

$$\mathrm{I}(X_1, X_2 \colon Y) = 2$$
$$\mathrm{I}(X_1 \colon Y) = 1$$
$$\mathrm{I}(X_2 \colon Y) = 1$$
$$\mathrm{I}_{\min}(\{X_1, X_2\} \colon Y) = 1$$
$$\mathrm{I}_\wedge(\{X_1, X_2\} \colon Y) = 1$$
$$\mathrm{I}_\alpha(\{X_1, X_2\} : Y) = 1$$



(b) circuit diagram



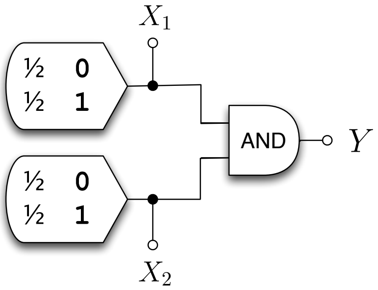
(c) WMS **(d)** $\mathrm{I}_{\min}$/$\mathrm{I}_\wedge$/I_α

Figure 3. Example RDNXOR. This is the canonical example of redundancy and synergy coexisting. $\mathrm{I}_{\min}$ and $\mathrm{I}_\wedge$ each reach the desired decomposition of one bit of redundancy and one bit of synergy. This example demonstrates $\mathrm{I}_\wedge$ correctly extracting the embedded redundant bit within X_1 and X_2.

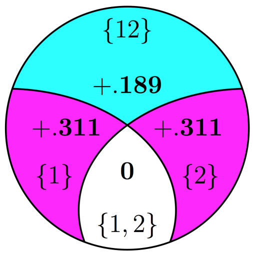
X_1	X_2	Y	
0	0	0	1/4
0	1	0	1/4
1	0	0	1/4
1	1	1	1/4

(a)
$\Pr(x_1, x_2, y)$

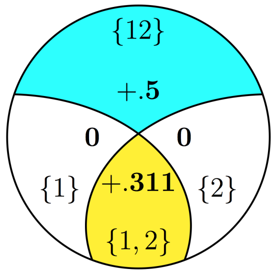
$$\mathrm{I}(X_1, X_2 : Y) = 0.811$$
$$\mathrm{I}(X_1 : Y) = 0.311$$
$$\mathrm{I}(X_2 : Y) = 0.311$$
$$\mathrm{I}_{\min}(\{X_1, X_2\} : Y) = 0.311$$
$$\mathrm{I}_{\wedge}(\{X_1, X_2\} : Y) = 0$$
$$\mathrm{I}_{\alpha}(\{X_1, X_2\} : Y) = 0$$



(b) circuit diagram



(c)
WMS/$\mathrm{I}_{\wedge}$/I_{α}



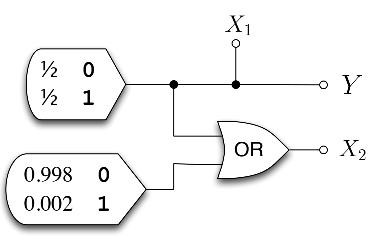
(d) $\mathrm{I}_{\min}$

Figure 4. Example AND. It is universally agreed that the redundant information is between $[0, 0.311]$ bits. The most compelling argument is from [15] arguing for 0.311 bits of redundant information.

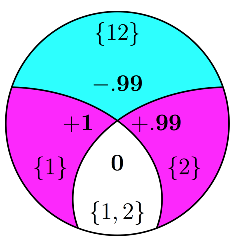
X_1	X_2	Y	
0	0	0	0.499
0	1	0	0.001
1	1	1	0.500

(a) $\Pr(x_1, x_2, y)$

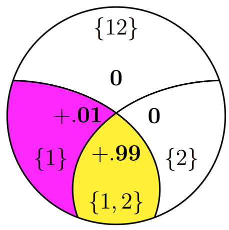
$$\mathrm{I}(X_1, X_2 : Y) = 1$$
$$\mathrm{I}(X_1 : Y) = 1$$
$$\mathrm{I}(X_2 : Y) = 0.99$$



(b) circuit diagram



(c) $\mathrm{I}_{\wedge}$



(d)
WMS/$\mathrm{I}_{\min}$/I_{α}

Figure 5. Example IMPERFECTRDN. $\mathrm{I}_{\wedge}$ is blind to the noisy correlation between X_1 and X_2 and calculates zero redundant information. An ideal $\mathrm{I}_{\cap}$ measure would detect that all of the information X_2 specifies about Y is also specified by X_1 to calculate $\mathrm{I}_{\cap}(\{X_1, X_2\} : Y) = 0.99$ bits.

4. Previous candidate measures

In [13], the authors propose to use the following quantity, $\mathrm{I}_{\min}$, as the intersection information measure:

$$\begin{aligned}\mathrm{I}_{\min}\left(X_1,\ldots,X_n:Y\right) &\equiv \sum_{y\in Y}\Pr(y)\min_{i\in\{1,\ldots,n\}}\mathrm{I}(X_i\!:\!Y=y)\\ &= \sum_{y\in Y}\Pr(y)\min_{i\in\{1,\ldots,n\}}\mathrm{D_{KL}}[\Pr(X_i|y)\|\Pr(X_i)]\ ,\end{aligned} \tag{3}$$

where $\mathrm{D_{KL}}$ is the Kullback-Leibler divergence.

Though $\mathrm{I}_{\min}$ is an intuitive and plausible choice for the intersection information, [9] showed that $\mathrm{I}_{\min}$ has counterintuitive properties. In particular, $\mathrm{I}_{\min}$ calculates one bit of redundant information for example UNQ (Figure 2). It does this because each input shares one bit of information with the output. However, it is quite clear that the shared informations are, in fact, different: X_1 provides the low bit, while X_2 provides the high bit. This led to the conclusion that $\mathrm{I}_{\min}$ *overestimates* the ideal intersection information measure by focusing only on *how much* information the inputs provide to the output. Another way to understand why $\mathrm{I}_{\min}$ overestimates redundancy in example UNQ is to imagine a hypothetical example where there are exactly two bits of unique information for every state $y\in Y$ and no synergy or redundancy. $\mathrm{I}_{\min}$ would calculate the redundancy as the minimum over both predictors which would be $\min[1,1]=1$ bit. Therefore $\mathrm{I}_{\min}$ would calculate 1 bit of redundancy even though by definition there was no redundancy but merely two bits of unique information.

Another candidate measure of synergy, WholeMinusSum (WMS) [9,16], calculates zero synergy and redundancy for Example RDNXOR, as opposed to the intuitive value of one bit of redundancy and one bit of synergy.

5. New candidate measures

5.1. The $\mathrm{I}_\wedge$ measure

Based on [17], we can consider a candidate intersection information as the maximum mutual information $\mathrm{I}(Q\!:\!Y)$ that some random variable Q conveys about Y, subject to Q being a function of each predictor $X_1,\ldots,X_n$. After some algebra, this leads to,

$$\begin{aligned}\mathrm{I}_\wedge(\{X_1,\ldots,X_n\}\!:\!Y) \equiv &\max_{\Pr(Q|Y)}\mathrm{I}(Q\!:\!Y)\\ &\text{subject to } \forall i\in\{1,\ldots,n\}:\mathrm{H}(Q|X_i)=0\end{aligned}\ , \tag{4}$$

which reduces to a simple expression in [12].

Example IMPERFECTRDN highlights the foremost shortcoming of $\mathrm{I}_\wedge$; $\mathrm{I}_\wedge$ does not detect "imperfect" or "lossy" correlations between X_1 and X_2. Instead, $\mathrm{I}_\wedge$ calculates zero redundant information, that $\mathrm{I}_\cap(\{X_1,X_2\}\!:\!Y) = 0$ bits. This arises from $\Pr(X_1=1,X_2=0) > 0$. If this were zero, IMPERFECTRDN reverts to being determined by the properties $(\mathbf{SR})$ and the $(\mathbf{M_0})$ equality condition. Due to the nature of the common random variable, $\mathrm{I}_\wedge$ only sees the "deterministic" correlations between X_1 and X_2—add even an iota of noise between X_1 and X_2 and $\mathrm{I}_\wedge$ plummets to zero. This highlights a related issue with $\mathrm{I}_\wedge$; it is not continuous—an arbitrarily small change in the probability distribution can result in a discontinuous jump in the value of $\mathrm{I}_\wedge$.

Despite this, $\mathrm{I}_\wedge$ is a useful stepping-stone, it captures what is inarguably redundant information (the common random variable). In addition, unlike earlier measures, $\mathrm{I}_\wedge$ satisfies $(\mathbf{TM})$.

5.2. The I_α *measure* Intuitively, we expect that if Q only specifies redundant information, that conditioning on any predictor X_i would vanquish all of the information Q conveys about Y. We take this intuition to its final conclusion and find it yields a tigther lowerbound on $\mathrm{I}_\cap$ than $\mathrm{I}_\wedge$. Moreover, I_α pleasantly reduces to a $\mathrm{I}_\wedge$ but loosens the constraint in Equation (4) from $\mathrm{H}(Q|X_i) = 0$ to $\mathrm{H}(Q|X_i) = \mathrm{H}(Q_i|X_i, Y)$:

$$\begin{aligned} \mathrm{I}_\alpha(\{X_1, \ldots, X_n\} : Y) \equiv & \max_{\Pr(Q|Y)} \mathrm{I}(Q : Y) \\ & \text{subject to } \forall i \in \{1, \ldots, n\} : \mathrm{I}(Q, X_i : Y) = \mathrm{I}(X_i : Y) \end{aligned} \tag{5}$$

$$\begin{aligned} = & \max_{\Pr(Q|Y)} \mathrm{I}(Q : Y) \\ & \text{subject to } \forall i \in \{1, \ldots, n\} : \mathrm{H}(Q|X_i) = \mathrm{H}(Q|X_i, Y) \end{aligned} . \tag{6}$$

This measure obtains the desired values for the canonical examples in Section 3. However, its implicit definition makes it more difficult to verify whether or not it satisfies the desired properties in Section 3. Pleasingly, I_α also satisfies $(\mathbf{TM})$. We can also show (See Lemmas 1 and 2 in Appendix A) that

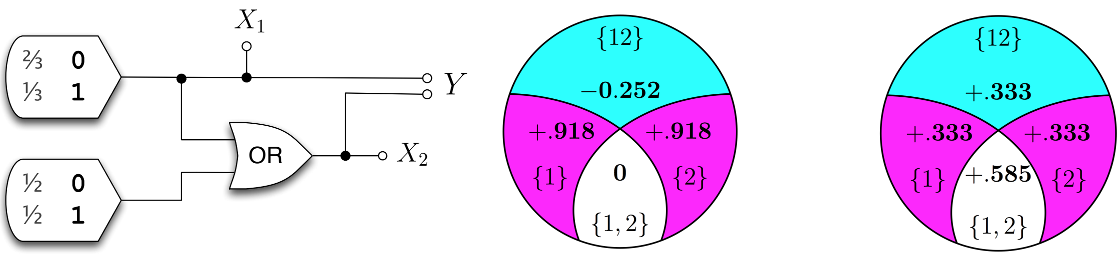
$$0 \le \mathrm{I}_\wedge(\{X_1, \ldots, X_n\} : Y) \le \mathrm{I}_\alpha(\{X_1, \ldots, X_n\} : Y) \le \mathrm{I}_{\min}(\{X_1, \ldots, X_n\} : Y) \ . \tag{7}$$

While I_α satisfies previously defined canonical examples, we have found another example, shown in Figure 6, for which $\mathrm{I}_\wedge$ and I_α both calculate negative synergy. This example further complicates Example AND by making the predictors mutually dependent.

X_1	X_2	Y	
0	0	00	1/3
0	1	01	1/3
1	1	11	1/3

(a) $\Pr(x_1, x_2, y)$

$$\begin{aligned} \mathrm{I}(X_1, X_2 : Y) &= 1.585 \\ \mathrm{I}(X_1 : Y) &= 0.918 \\ \mathrm{I}(X_2 : Y) &= 0.918 \\ \mathrm{I}(X_1 : X_2) &= 0.252 \end{aligned}$$



(b) circuit diagram **(c)** $\mathrm{I}_\wedge/\mathrm{I}_\alpha$ **(d)** $\mathrm{I}_{\min}$

Figure 6. Example SUBTLE.

6. Conclusion

The important part of this paper is exchanging $(\mathbf{M_0})$ with $(\mathbf{M_1})$ thus further constraining the space of acceptable $\mathrm{I}_\cap$ measures. The complexity community aspires to eventually find a unique $\mathrm{I}_\cap$ measure that satisfies a large portion of the desired properties, and *any noncontroversial tightening of the space of possible* $\mathrm{I}_\cap$ *measures, even (or especially?) if obvious in hindsight, is immensely welcome*.

As discussed in [12], $\mathrm{I}_\cap$ measures fail $(\mathbf{LP})$ if and only if they are *too strict* a measure of redundant information. Loosening the constraints on $\mathrm{I}_\wedge$ yields I_α and achieves a nonnegative decomposition on example IMPERFECTRDN. A natural next step is to loosen the constraints on I_α until achieving a nonnegative decomposition for example SUBTLE. Alternatively, a very plausible measure of the "unique information" [9,15,18] that satisfies $(\mathbf{LP})$ for $n = 2$ yet does not satisfy $(\mathbf{TM})$. It seems that $(\mathbf{LP})$ and $(\mathbf{TM})$ will be incompatible, and it would be nice to prove this.

Acknowledgments

We thank Jim Beck, Yaser Abu-Mostafa, Edwin Chong, Chris Ellison, and Ryan James for helpful discussions.

Author Contributions

Both authors shared in this research equally. Both authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

A. Appendix

Proof I_α **does not satisfy** $(\mathbf{LP})$. Proof by counter-example SUBTLE (Figure 6).

For $\mathrm{I}(Q\!:\!Y|X_1) = 0$, then Q must not distinguish between states of $Y = \texttt{00}$ and $Y = \texttt{01}$ (because X_1 does not distinguish between these two states). This entails that $\Pr(Q|Y = \texttt{00}) = \Pr(Q|Y = \texttt{01})$. By symmetry, likewise for $\mathrm{I}(Q\!:\!Y|X_2) = 0$, Q must be distinguish between states $Y = \texttt{01}$ and $Y = \texttt{11}$. Altogether, this entails that $\Pr(Q|Y = \texttt{00}) = \Pr(Q|Y = \texttt{01}) = \Pr(Q|Y = \texttt{11})$, which then entails, $\Pr(q|y_i) = \Pr(q|y_j) \quad \forall q \in Q, y_i \in Y, y_j \in Y$, which is only achievable when $\Pr(q) = \Pr(q|y) \quad \forall q \in Q, y \in Y$. This makes $\mathrm{I}(Q\!:\!Y) = 0$, therefore for example SUBTLE, $\mathrm{I}_\alpha(\{X_1, X_2\} : Y) = 0$.

Lemma 1. *We have* $\mathrm{I}_\wedge(\{X_1, \ldots, X_n\}\!:\!Y) \leq \mathrm{I}_\alpha(\{X_1, \ldots, X_n\} : Y)$.

Proof. We define a random variable $Q' = X_1 \wedge \cdots \wedge X_n$. We then plugin Q' for Q in the definition of I_α. This newly plugged-in Q satisfies the constraint $\forall i \in \{1, \ldots, n\}$ that $\mathrm{I}(Q\!:\!Y|X_i) = 0$. Therefore, Q' is always a possible choice for Q, and the maximization of $\mathrm{I}(Q\!:\!Y)$ in I_α must be at least as large as $\mathrm{I}(Q'\!:\!Y) = \mathrm{I}_\wedge(\{X_1, \ldots, X_n\}\!:\!Y)$. $\square$

Lemma 2. *We have* $\mathrm{I}_\alpha(\{X_1, \ldots, X_n\} : Y) \leq \mathrm{I}_{\min}(X_1, \ldots, X_n : Y)$

Proof. For a given state $y \in Y$ and two arbitrary random variables Q and X, given $\mathrm{I}(Q\!:\!y|X) = \mathrm{D}_{\mathrm{KL}}[\Pr(QX|y) \| \Pr(Q|X)\Pr(X|y)] = 0$, we show that, $\mathrm{I}(Q\!:\!y) \leq \mathrm{I}(X\!:\!y)$,

$$\begin{aligned}\mathrm{I}(X\!:\!y) - \mathrm{I}(Q\!:\!y) &= \sum_{x \in X} \Pr(x|y) \log \frac{\Pr(x|y)}{\Pr(x)} - \sum_{q \in Q} \Pr(q|y) \log \frac{\Pr(q|y)}{\Pr(q)} \\ &\geq 0 \,.\end{aligned}$$

Generalizing to n predictors $X_1, \ldots, X_n$, the above shows that that the maximum $\mathrm{I}(Q\!:\!y)$ under constraint $\mathrm{I}(Q\!:\!y|X_i)$ will always be less than $\min_{i \in \{1,\ldots,n\}} \mathrm{I}(X_i\!:\!y)$, which completes the proof. □

Lemma 3. *Measure* $\mathrm{I}_{\min}$ *satisfies desired property Strong Monotonicity,* $(\mathbf{M_1})$.

Proof. Given $\mathrm{H}(Y|Z) = 0$, then the specific-surprise $\mathrm{I}(Z\!:\!y)$ yields,

$$\begin{aligned}\mathrm{I}(Z\!:\!y) &\equiv \mathrm{D}_{\mathrm{KL}}[\Pr(Z|y) \| \Pr(Z)] \\ &= \sum_{z \in Z} \Pr(z|y) \log \frac{\Pr(z|y)}{\Pr(z)} \\ &= \sum_{z \in Z} \Pr(z|y) \log \frac{1}{\Pr(y)} \\ &= \log \frac{1}{\Pr(y)} \,.\end{aligned}$$

Given that for an arbitrary random variable X_i, $\mathrm{I}(X_i\!:\!y) \leq \log \frac{1}{\Pr(y)}$. As $\mathrm{I}_{\min}$ takes only uses the $\min_i \mathrm{I}(X_i\!:\!y)$, the minimum is invariant under adding any predictor Z such that $\mathrm{H}(Y|Z) = 0$. Therefore, measure $\mathrm{I}_{\min}$ satisfies property $(\mathbf{M_1})$. □

References

1. Schneidman, E.; Bialek, W.; Berry, M.J. Synergy, redundancy, and independence in population codes. *J. Neurosci.* **2003**, *23*, 11539–11553.
2. Narayanan, N.S.; Kimchi, E.Y.; Laubach, M. Redundancy and Synergy of Neuronal Ensembles in Motor Cortex. *J. Neurosci.* **2005**, *25*, 4207–4216.
3. Balduzzi, D.; Tononi, G. Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Comput. Biol.* **2008**, *4*, e1000091.
4. Anastassiou, D. Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* **2007**, *3*, 83.
5. Lizier, J.T.; Flecker, B.; Williams, P.L. Towards a Synergy-based Approach to Measuring Information Modification. In Proceedings of 2013 IEEE Symposium on Artificial Life (ALIFE), Singapore, Singapore, 16–19 April 2013; pp. 43–51.
6. Gács, P.; Körner, J. Common information is far less than mutual information. *Prob. Control Inf. Theory* **1973**, *2*, 149–162.
7. Wyner, A.D. The common information of two dependent random variables. *IEEE Trans. Inf. Theory* **1975**, *21*, 163–179.

8. Kumar, G.R.; Li, C.T.; Gamal, A.E. Exact Common Information. In Proceedings of 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 161–165.
9. Griffith, V.; Koch, C. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*; Emergence, Complexity and Computation Serie, Volume 9; Springer: Berlin/Heidelberg, Germany, 2014; pp. 159–190.
10. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130.
11. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared Information—New Insights and Problems in Decomposing Information in Complex Systems. In *Proceedings of European Conference on Complex Systems 2012*; Springer Proceedings in Complexity Serie; Springer: Switzerland, 2013; pp. 251–269.
12. Griffith, V.; Chong, E.K.P.; James, R.G.; Ellison, C.J.; Crutchfield, J.P. Intersection Information based on Common Randomness. *Entropy* **2014**, *16*, 1985–2000.
13. Williams, P.L.; Beer, R.D. Nonnegative Decomposition of Multivariate Information. **2010**, arXiv:1004.2515.
14. Weisstein, E.W. Antichain. Available online: http://mathworld.wolfram.com/Antichain.html (accessed on 29 June 2015).
15. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183.
16. Schneidman, E.; Still, S.; Berry, M.J.; Bialek, W. Network Information and Connected Correlations. *Phys. Rev. Lett.* **2003**, *91*, 238701–238705.
17. Wolf, S.; Wullschleger, J. Zero-error information and applications in cryptography. In Proceedings of IEEE Information Theory Workshop, San Antonio, TX, USA, 24–29 October 2004; pp. 1–6.
18. Rauh, J.; Bertschinger, N.; Olbrich, E.; Jost, J. Reconsidering unique information: Towards a multivariate information decomposition. In Proceedings of 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 2232–2236.