Abstract

Since the seminal paper [1], there’s been a flurry of research [2–5] towards defining an intersection information that quantifies how much of “the same information” two or more random variables specify about a target random variable. A palatable measure of intersection information would provide a principled way to quantify slippery concepts such as synergy. Here we introduce an intersection information measure based on the Gács-Körner common random variable which is the first to satisfy the coveted Target Monotonicity property. Our measure is imperfect and we suggest directions for improvement.

1 Introduction

Introduced in [1], Partial Information Decomposition (PID) is an immensely useful framework for deepening our understanding of multivariate interactions—particularly our understanding of informational redundancy and synergy. To harness the PID framework, the user brings her own measure of intersection information, $I_{\cap}(X_1, \ldots, X_n : Y)$, which quantifies the magnitude of information that each of the $n$ predictors $X_1, \ldots, X_n$ conveys about a target random variable $Y$. An antichain lattice of redundant, unique, and synergistic partial informations is built from the intersection information.

In [1], the authors propose to use the following quantity, $I_{\min}$, as the intersection information measure:

\[
I_{\min}(X_1, \ldots, X_n : Y) \equiv \sum_y \Pr(y) \min_i I(X_i : Y = y) \\
= \sum_y \Pr(y) \min_i D_{KL}\left[\Pr(X_i | y) \left\| \Pr(X_i)\right]\right],
\]

where $D_{KL}$ is the Kullback-Leibler divergence.
Though \( I_{\text{min}} \) is an intuitive and plausible choice for the intersection information, [2] showed that \( I_{\text{min}} \) has counterintuitive properties. In particular, \( I_{\text{min}} \) calculates one bit of redundant information for example \( \text{UNQ} \) (Figure 1). It does this because each input shares one bit of information with the output. However, its 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 \( I_{\text{min}} \) over-estimates the ideal intersection information measure by focusing only on how much information the inputs provide to the output. An ideal measure of intersection information must recognize that there are non-equivalent ways of providing information to the output. The search for an improved intersection information measure ensued, continued through [3–5], and today a widely accepted intersection information measure remains undiscovered.

Here we do not definitively solve this problem, but we present a strong candidate intersection information measure for the special case of zero-error information. This is useful in of itself because it provides a template for how the yet undiscovered ideal intersection information measure for Shannon mutual information could work. Alternatively, if a Shannon intersection information measure with the same properties does not exist, then we have learned something significant.

In the next section, we introduce some definitions, some notation, and a necessary lemma. We also extend and clarify the desired properties for intersection information. In Section 3 we introduce zero-error information and its intersection information measure. In Section 4 we use the same methodology to produce a novel candidate for the Shannon intersection information. In Section 5 we show the successes and shortcomings of our candidate intersection information measure using example circuits. Finally in Section 7, we summarize our progress towards the ideal intersection information measure and suggest directions for improvement. The Appendix is devoted to technical lemmas and their proofs, to which we refer in the main text.

2 Preliminaries

2.1 Informational Partial Order and Equivalence

We assume an underlying probability space on which we define random variables denoted by capital letters (e.g., \( X \), \( Y \), and \( Z \)). In this paper, we consider only random variables taking values on finite spaces.

Given random variables \( X \) and \( Y \), we write \( X \preceq Y \) to signify that there exists a measurable function \( f \) such that \( X = f(Y) \) almost surely (i.e., with probability one). In this case, following the terminology in [6], we say that \( X \) is informationally poorer than \( Y \); this induces a partial order on the set of random variables. Similarly, we write \( X \succeq Y \) if \( Y \preceq X \), in which case we say \( X \) is informationally richer than \( Y \).

If \( X \) and \( Y \) are such that \( X \preceq Y \) and \( X \succeq Y \), then we write \( X \equiv Y \). In this case, again following [6], we say that \( X \) and \( Y \) are informationally equivalent. In other words, \( X \equiv Y \) if and only if one can relabel the values of \( X \) to obtain a random value that is equal to \( Y \) almost surely, and vice versa.

This “information-equivalence” relation can easily be shown to be an equivalence relation, so that we can partition the set of all random variables into disjoint equivalence classes. The \( \preceq \) ordering is invariant within these equivalence classes in the following sense. If \( X \preceq Y \) and \( Y \cong Z \), then \( X \preceq Z \). Similarly, if \( X \preceq Y \) and \( X \cong Z \), then \( Z \preceq Y \). Moreover, within each equivalence class, the entropy is invariant, as stated formally in Lemma 1 below.

2.2 Information Lattice

Next, we follow [6] and consider the join and meet operators. These operators were defined for information elements, which are \( \sigma \)-algebras, or, equivalently, equivalence classes of random variables. We deviate from [6], though, by defining the join and meet operators for random variables, but we do preserve their conceptual properties.
Given random variables $X$ and $Y$, we define $X \g Y$ (called the join of $X$ and $Y$) to be an informationally poorest (“smallest” in the sense of the partial order $\preceq$) random variable such that $X \preceq X \g Y$ and $Y \preceq X \g Y$. In other words, if $Z$ is such that $X \preceq Z$ and $Y \preceq Z$, then $X \g Y \preceq Z$. Note that $X \g Y$ is unique only up to equivalence with respect to $\approx$. In other words, $X \g Y$ does not define a specific, unique random variable. Nonetheless, standard information-theoretic quantities are invariant over the set of random variables satisfying the condition specified above. For example, the entropy of $X \g Y$ is invariant over the entire equivalence class of random variables satisfying the condition above (by Lemma 1(a) below). Similarly, the inequality $Z \preceq X \g Y$ does not depend on the specific random variable chosen, as long as it satisfies the condition above. Note that the pair $(X, Y)$ is an instance of $X \g Y$.

In a similar vein, given random variables $X$ and $Y$, we define $X \& Y$ (called the meet of $X$ and $Y$) to be an informationally richest random variable (“largest” in the sense of $\succeq$) such that $X \& Y \succeq X$ and $X \& Y \succeq Y$. In other words, if $Z$ is such that $Z \succeq X$ and $Z \succeq Y$, then $Z \succeq X \& Y$. Following [7], we also call $X \& Y$ the common random variable of $X$ and $Y$. Again, considering the entropy of $X \& Y$ or the inequality $Z \succeq X \& Y$ does not depend on the specific random variable chosen, as long as it satisfies the condition above.

The $\g$ and $\&$ operators satisfy the algebraic properties of a lattice [6]. In particular, the following hold:

- commutative laws: $X \g Y \equiv Y \g X$ and $X \& Y \equiv Y \& X$
- associative laws: $X \g (Y \g Z) \equiv (X \g Y) \g Z$ and $X \& (Y \& Z) \equiv (X \& Y) \& Z$
- absorption laws: $X \g (X \& Y) \equiv X$ and $X \& (X \g Y) \equiv X$
- idempotent laws: $X \g X \equiv X$ and $X \& X \equiv X$
- generalized absorption laws: if $X \preceq Y$, then $X \g Y \equiv Y$ and $X \& Y \equiv X$.

Finally, the partial order $\preceq$ is preserved under $\g$ and $\&$, i.e., if $X \preceq Y$, then $X \g Z \preceq Y \g Z$ and $X \& Z \preceq X \& Z$.

### 2.3 Invariance and Monotonicity of Entropy

Let $H(\cdot)$ represent the entropy function, and $H(\cdot|\cdot)$ the conditional entropy. To be consistent with the colon in the intersection information, we denote the Shannon mutual information between $X$ and $Y$ by $I(X:Y)$ instead of the more common $I(X; Y)$. Lemma 1 establishes the invariance and monotonicity of the entropy and conditional entropy functions with respect to $\equiv$ and $\preceq$.

**Lemma 1.** The following hold:

(a) If $X \equiv Y$, then $H(X) = H(Y)$, $H(X|Z) = H(Y|Z)$, and $H(Z|X) = H(Z|Y)$.

(b) If $X \preceq Y$, then $H(X) \leq H(Y)$, $H(X|Z) \leq H(Y|Z)$, and $H(Z|X) \geq H(Z|Y)$.

(c) $X \preceq Y$ if and only if $H(X|Y) = 0$.

**Proof.** Part (a) follows from [6], Proposition 1. Part (c) follows from [6], Proposition 4. The first two desired inequalities in part (b) follow from [6], Proposition 5. Now we show that if $X \preceq Y$, then $H(Z|X) \geq H(Z|Y)$. Suppose that $X \preceq Y$. Then, by the generalized absorption law, $X \g Y \equiv Y$. We have

\[
I(Z:Y) = H(Y) - H(Y|Z)
= H(Z \g Y) - H(Z \g Y|Z) \quad \text{by part (a)}
= I(Z:Z \g Y)
= I(Z:X) + I(Z:Y|X)
\geq I(Z:X).
\]

Substituting $I(Z:Y) = H(Z) - H(Z|Y)$ and $I(Z:X) = H(Z) - H(Z|X)$, we obtain $H(Z|X) \geq H(Z|Y)$ as desired. \qed
Remark: Because \((X, Y) \cong X \equiv Y\) as noted before, we also have \(H(X, Y) = H(X \equiv Y)\) by Lemma 1(a).

2.4 Desired Properties of Intersection Information

There are currently 12 intuitive properties that we wish the ideal intersection information measure \(I_\gamma\) to satisfy. Some are new (e.g. \((M_1), (Eq), (LB)\)), but most were introduced earlier, in various forms, Refs. [1–5]. They are as follows:

\((GP)\) Global Positivity: \(I_\gamma(X_1, \ldots, X_n: Y) \geq 0\), and \(I_\gamma(X_1, \ldots, X_n: Y) = 0\) if \(Y\) is a constant.

\((Eq)\) Equivalence-Class Invariance: \(I_\gamma(X_1, \ldots, X_n: Y)\) is invariant under substitution of \(X_i\) (for any \(i = 1, \ldots, n\)) or \(Y\) by an informationally equivalent random variable.

\((TM)\) Target Monotonicity: If \(Y \preceq Z\), then \(I_\gamma(X_1, \ldots, X_n: Y) \leq I_\gamma(X_1, \ldots, X_n: Z)\).

\((M_0)\) Weak Monotonicity: \(I_\gamma(X_1, \ldots, X_n, W: Y) \leq I_\gamma(X_1, \ldots, X_n: Y)\) with equality if there exists \(Z \in \{X_1, \ldots, X_n\}\) such that \(Z \preceq W\).

\((S_0)\) Weak Symmetry: \(I_\gamma(X_1, \ldots, X_n: Y)\) is invariant under reordering of \(X_1, \ldots, X_n\).

Remark: If \((S_0)\) is satisfied, the first argument of \(I_\gamma(X_1, \ldots, X_n: Y)\) can be treated as a set of random variables rather than a list. In this case, the notation \(I_\gamma(\{X_1, \ldots, X_n\}: Y)\) would also be appropriate.

For the next set of properties, \(\mathcal{I}(X: Y)\) is a given normative measure of information between \(X\) and \(Y\). For example, \(\mathcal{I}(X: Y)\) could denote the Shannon mutual information; i.e., \(\mathcal{I}(X: Y) = I(X: Y)\). Alternatively, as discussed in the next section, we might take \(\mathcal{I}(X: Y)\) to be the zero-error information. Yet other possibilities for \(\mathcal{I}(X: Y)\) include the Wyner common information [8] or the quantum mutual information [9]. The following are desired properties of intersection information relative to the given information measure \(\mathcal{I}\).

\((LB)\) Lowerbound: If \(Q \preceq X_i\) for all \(i = 1, \ldots, n\), then \(I_\gamma(X_1, \ldots, X_n: Y) \geq \mathcal{I}(Q: Y)\). Under a mild assumption,\(^3\) this equates to \(I_\gamma(X_1, \ldots, X_n: Y) \geq \mathcal{I}(X_1 \wedge \cdots \wedge X_n : Y)\).

\((SR)\) Self-Redundancy: \(I_\gamma(X_1: Y) = \mathcal{I}(X_1: Y)\). The intersection information a single predictor \(X_1\) conveys about the target \(Y\) is equal to the information between the predictor and the target given by the information measure \(\mathcal{I}\).

\((Id)\) Identity: \(I_\gamma(X, Y: X \equiv Y) = \mathcal{I}(X: Y)\).

\((LP_0)\) Weak Local Positivity: \(I_\gamma(X_1, X_2: Y) \geq \mathcal{I}(X_1: Y) + \mathcal{I}(X_2: Y) - \mathcal{I}(X_1 \equiv X_2: Y)\). In other words, for \(n = 2\) predictors, the derived “partial informations” defined in [1] are nonnegative when both \((LP_0)\) and \((GP)\) hold.

Finally, we have the less obvious “strong” properties.

\((M_1)\) Strong Monotonicity: \(I_\gamma(X_1, \ldots, X_n, W: Y) \leq I_\gamma(X_1, \ldots, X_n: Y)\) with equality if there exists \(Z \in \{X_1, \ldots, X_n\}\) such that \(Z \preceq W\).

\((S_1)\) Strong Symmetry: \(I_\gamma(X_1, \ldots, X_n: Y)\) is invariant under reordering of \(X_1, \ldots, X_n, Y\).

\((LP_1)\) Strong Local Positivity: For all \(n\), the derived “partial informations” defined in [1] are nonnegative.

Properties \((Eq), (LB),\) and \((M_1)\) are novel and are introduced for the first time here. Given \(I_\gamma, X_1, \ldots, X_n, Y,\) and \(Z\), we define the conditional \(I_\gamma\) as:

\[
I_\gamma(X_1, \ldots, X_n: Z|Y) = I_\gamma(X_1, \ldots, X_n: Y \equiv Z) - I_\gamma(X_1, \ldots, X_n: Y) .
\]

This definition of \(I_\gamma(X_1, \ldots, X_n: Z|Y)\) gives rise to the familiar “chain rule”:

\[
I_\gamma(X_1, \ldots, X_n: Y \equiv Z) = I_\gamma(X_1, \ldots, X_n: Y) + I_\gamma(X_1, \ldots, X_n: Z|Y) .
\]

Some provable\(^2\) properties are:

\(^1\)See Lemmas 2 and 3 in Appendix C.1.

\(^2\)See Lemma 4 in Appendix C.1.
\[ I_\perp(X_1, \ldots, X_n : Z | Y) \geq 0. \]
\[ I_\perp(X_1, \ldots, X_n : Z | Y) = I_\perp(X_1, \ldots, X_n : Z) \] if \( Y \) is a constant.

### 3 Candidate Intersection Information for Zero-Error Information

#### 3.1 Zero-Error Information

Introduced in [10], the zero-error information, or Gács-Körner common information, is a stricter variant of Shannon mutual information. Whereas the mutual information \( I(A : B) \) quantifies the magnitude of information \( A \) conveys about \( B \) with an arbitrarily small error \( \epsilon > 0 \), the zero-error information, denoted \( I_0^0(A : B) \), quantifies the magnitude of information \( A \) conveys about \( B \) with exactly zero error, i.e., \( \epsilon = 0 \). The zero-error information between \( A \) and \( B \) equals the entropy of the common random variable \( A \perp B \),
\[ I_0^0(A : B) \equiv H(A \perp B). \]

An algorithm for computing an instance of the common random variable between two random variables is provided in [10], and straightforwardly generalizes to \( n \) random variables.\(^3\)

Zero-error information has several notable properties, but the most salient is that it is nonnegative and bounded by the mutual information,
\[ 0 \leq I_0^0(A : B) \leq I(A : B). \]

This generalizes to arbitrary \( n \):
\[ 0 \leq I_0^0(X_1 : \cdots : X_n) \leq \min_{i,j} I(X_i : X_j). \]

#### 3.2 Intersection Information for Zero-Error Information

It is pleasingly straightforward to define a palatable intersection information for zero-error information (i.e., setting \( I = I_0^0 \) as the normative measure of information). We propose the zero-error intersection information, \( I_\perp^0(X_1, \ldots, X_n : Y) \), as the maximum zero-error information \( I_0^0(Q : Y) \) that some random variable \( Q \) conveys about \( Y \), subject to \( Q \) being a function of each predictor \( X_1, \ldots, X_n \):
\[ I_\perp^0(X_1, \ldots, X_n : Y) \equiv \max_{P_r(Q | Y)} I_0^0(Q : Y) \]
subject to \( Q \perp X_i \forall i \in \{1, \ldots, n\} \).

Basic algebra\(^4\) shows that a maximizing \( Q \) is the common random variable across all predictors. This substantially simplifies eq. (2) to:
\[ I_\perp^0(X_1, \ldots, X_n : Y) = I_0^0(X_1 \perp \cdots \perp X_n : Y) \]
\[ = H[(X_1 \perp \cdots \perp X_n) \perp Y] \]
\[ = H(X_1 \perp \cdots \perp X_n \perp Y). \]

Importantly, the zero-error information, \( I_0^0(X_1, \ldots, X_n ; Y) \), satisfies 10 of the 12 desired properties from Section 2.4, leaving only \( (LP_0) \) and \( (LP_1) \) unsatisfied.\(^5\)

### 4 Candidate Intersection Information for Shannon Information

In the last section, we defined an intersection information for zero-error information which satisfies the vast majority of desired properties. This is a solid start, but an intersection information for Shannon mutual information remains the goal. Towards this end, we use the

---

\(^3\)See Appendix A.

\(^4\)See Lemma 11 in Appendix D.

\(^5\)See Lemmas 5, 6, 7 in Appendix C.2.
same method as in eq. (2), leading to $I_{\lambda}$, our candidate intersection information measure for Shannon mutual information,

$$I_{\lambda}(X_1, \ldots, X_n : Y) \equiv \max_{Q(Y)} I(Q : Y)$$

subject to $Q \preceq X_i \forall i \in \{1, \ldots, n\}.$

(4)

With some algebra\(^6\) this similarly simplifies to,

$$I_{\lambda}(X_1, \ldots, X_n : Y) = I(X_1 \wedge \cdots \wedge X_n : Y).$$

(5)

Unfortunately $I_{\lambda}$ does not satisfy as many of the desired properties as $I^0_{\lambda}$. However, our candidate $I_{\lambda}$ still satisfies 7 of the 12 properties—most importantly the enviable (TM),\(^7\) which has, until now, not been satisfied by any proposed measure. Table 1 lists the desired properties satisfied by $I_{\min}$, $I_{\lambda}$, and $I^0_{\lambda}$. For reference, we also include $I_{\red}$, the proposed measure from [3].

Comparing the three subject intersection information measures,\(^8\) we have:

$$0 \leq I^0_{\lambda}(X_1, \ldots, X_n : Y) \leq I_{\lambda}(X_1, \ldots, X_n : Y) \leq I_{\min}(X_1, \ldots, X_n : Y).$$

(6)

<table>
<thead>
<tr>
<th>Property</th>
<th>$I_{\min}$</th>
<th>$I_{\red}$</th>
<th>$I_{\lambda}$</th>
<th>$I^0_{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(GP) Global Positivity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(Eq) Equivalence-Class Invariance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(TM) Target Monotonicity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(M_0) Weak Monotonicity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(S_0) Weak Symmetry</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(LB) Lowerbound</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(SR) Self-Redundancy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(Id) Identity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(LP_0) Weak Local Positivity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(M_1) Strong Monotonicity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(S_1) Strong Symmetry</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(LP_1) Strong Local Positivity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: The $I_{\lambda}$ desired properties each measure satisfies.

Despite not satisfying (LP_0), $I_{\lambda}$ remains an important stepping-stone towards the ideal Shannon $I_\lambda$. First, $I_{\lambda}$ captures what is inarguably redundant information (the common random variable)—this makes $I_{\lambda}$ necessarily a lower bound on any reasonable redundancy measure. Second, it is the first proposal to satisfy target monotonicity and the associated chain rule. Lastly, $I_{\lambda}$ is the first measure to reach intuitive answers in many canonical situations, while also being generalizable to an arbitrary number of inputs.

5 Three Examples Comparing $I_{\min}$ and $I_{\lambda}$

Examples Unq and RdnXor illustrate $I_{\lambda}$’s successes and example ImperfectRdn illustrates $I_{\lambda}$’s paramount deficiency. For each example we show the joint distribution $\Pr(x_1, x_2, y)$, a

\(^{6}\)See Lemma 12 in Appendix D.

\(^{7}\)See Lemmas 8, 9, 10 in Appendix C.3.

\(^{8}\)See Lemma 13 in Appendix D.
Therefore, as $I(X_1 : X_2) = 0$, we have $I_\lambda(X_1, X_2 : Y) = 0$ bits leading to $I_\partial(X_1 : Y) = 1$ bit and $I_\partial(X_2 : Y) = 1$ bit (Figure 1d).

**Example RdnXor** (Figure 2). In [2], RdnXor was an example where $I_{\min}$ shined by reaching the desired decomposition of one bit of redundancy and one bit of synergy. We see

---

9This is the same notation used in [4].
To summarize, IMPERFECTRDN shows that when there are additional “imperfect” correlations between $A$ and $B$, i.e. $I(A:B|A \lor B) > 0$, $I_\lambda$ sometimes underestimates the ideal $I_\nabla(A,B;Y)$. That $I_\lambda$ finds this same answer. $I_\lambda$ extracts the common random variable within $X_1$ and $X_2$, the $r/R$ bit, and calculates the mutual information between the common random variable and $Y$ to arrive at $I_\lambda(X_1,X_2;Y) = 1$ bit.

**Example ImperfectRdn** (Figure 3). IMPERFECTRDN highlights the foremost shortcoming of $I_\lambda$: $I_\lambda$ does not detect “imperfect” or “lossy” correlations between $X_1$ and $X_2$. Given $\text{LP}_0$, we can determine the desired decomposition analytically. First, $I(X_1 \cap X_2;Y) = I(X_1;Y) = 1$ bit; therefore, $I(X_2;Y|X_1) = I(X_1 \cap X_2;Y) - I(X_1;Y) = 0$ bits. This determines two of the partial informations—the synergistic information $I_{0}(X_1 \cap X_2;Y)$ and the unique information $I_{0}(X_2;Y)$ are both zero. Then, the redundant information $I_{0}(X_1,X_2;Y) = I(X_2;Y) - I_{0}(X_2;Y) = I(X_2;Y) = 0.99$ bits. Having determined three of the partial informations, we compute the final unique information $I_{0}(X_1;Y) = I(X_1;Y) - 0.99 = 0.01$ bits.

How well do $I_{\min}$ and $I_\lambda$ match the desired decomposition of IMPERFECTRDN? We see that $I_{\min}$ calculates the desired decomposition (Figure 3c); however, $I_\lambda$ does not (Figure 3d). Instead, $I_\lambda$ calculates zero redundant information, that $I_\nabla(X_1,X_2;Y) = 0$ bits. This unpleasant answer arises from $\Pr(X_1 = 1, X_2 = 0) > 0$. If this were zero, IMPERFECTRDN reverts to the example RDN (Figure 5 in Appendix E) where both $I_{\lambda}$ and $I_{\min}$ reach the desired one bit of redundant information. Due to the nature of the common random variable, $I_\lambda$ only sees the “deterministic” correlations between $X_1$ and $X_2$—add even an iota of noise between $X_1$ and $X_2$ and $I_\lambda$ plummets to zero. This highlights a related issue with $I_\lambda$: it is not continuous—an arbitrarily small change in the probability distribution can result in a discontinuous jump in the value of $I_\lambda$. As with traditional information measures such as the entropy and the mutual information, it may be desirable to have an $I_\nabla$ measure that is continuous over the simplex.

To summarize, IMPERFECTRDN shows that when there are additional “imperfect” correlations between $A$ and $B$, i.e. $I(A:B|A \lor B) > 0$, $I_\lambda$ sometimes underestimates the ideal $I_\nabla(A,B;Y)$. 

---

### Table: Pr($x_1, x_2, y$)

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$y$</th>
<th>$\Pr(x_1, x_2, y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_0$</td>
<td>$r_0$</td>
<td>$r_0$</td>
<td>1/8</td>
</tr>
<tr>
<td>$r_0$</td>
<td>$r_1$</td>
<td>$r_1$</td>
<td>1/8</td>
</tr>
<tr>
<td>$r_1$</td>
<td>$r_0$</td>
<td>$r_1$</td>
<td>1/8</td>
</tr>
<tr>
<td>$r_1$</td>
<td>$r_1$</td>
<td>$r_0$</td>
<td>1/8</td>
</tr>
</tbody>
</table>

---

### Figure 2: Example RDNXOR.

This is the canonical example of redundancy and synergy coexisting. $I_{\min}$ and $I_\lambda$ each reach the desired decomposition of one bit of redundancy and one bit of synergy. This is the simplest example demonstrating $I_\lambda$ and $I_\nabla$ correctly extracting the embedded redundant bit within $X_1$ and $X_2$.
<table>
<thead>
<tr>
<th>$X_1, X_2$</th>
<th>$Y$</th>
<th>$I(X_1 \land X_2 : Y) = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td>0</td>
<td>0.499</td>
</tr>
<tr>
<td>0 1</td>
<td>0</td>
<td>0.001</td>
</tr>
<tr>
<td>1 1</td>
<td>1</td>
<td>0.500</td>
</tr>
</tbody>
</table>

(a) $P_r(x_1, x_2, y)$

(b) Circuit diagram

(c) $I_{min}$

(d) $I_{\lambda}$

(e) $I_{\lambda}^\dagger$

Figure 3: Example IMPERFECTRDN. $I_{\lambda}$ is blind to the noisy correlation between $X_1$ and $X_2$ and calculates zero redundant information. An ideal $I_\cap$ measure would detect that all of the information $X_2$ specifies about $Y$ is also specified by $X_1$ to calculate $I_\cap(X_1, X_2 : Y) = 0.99$ bits.
6 Negative synergy and state-dependent (GP)

In IMPERFECTRDN we saw $I_\lambda$ calculate a synergy of $-0.99$ bits (Figure 3d). What does this mean? Could negative synergy be a “real” property of Shannon information? When $n = 2$, it’s fairly easy to diagnose the cause of negative synergy from the equation for $I_\theta(X_1, X_2 : Y)$ in eq. (7). Given (GP) and (SR), negative synergy occurs if and only if,

$$I(X_1 \cap X_2 : Y) < I(X_1 : Y) + I(X_2 : Y) - I_\cap(X_1, X_2 : Y) = I_\cup(X_1, X_2 : Y). \quad (9)$$

From eq. (9), we see negative synergy occurs when $I_\cap$ is small, perhaps too small. Equivalently, negative synergy occurs when the joint r.v. conveys less about $Y$ than the two r.v.’s $X_1$ and $X_2$ convey separately—mathematically, when $I(X_1 \cap X_2 : Y) < I_\cup(X_1, X_2 : Y).$ On the face of it this sounds strange. No usable structure in $X_1$ or $X_2$ “disappears” after they are combined by $Z = X_1 \cap X_2$. By the definition of $\cap$, there are always functions $f_1$ and $f_2$ such that $X_1 \equiv f_1(Z)$ and $X_2 \equiv f_2(Z)$. Therefore, if your favorite $I_\cap$ measure does not satisfy (LP$_0$), it is likely too strict.

This means that, to our surprise, our measure $I_\lambda$ does not account for the full zero-information overlap between $I_0(X_1 : Y)$ and $I_0(X_2 : Y)$. This is shown in example SUBTLE (Figure 4) where $I_\lambda$ calculates a synergy of $-0.252$ bits. Defining a zero-error $I_\cap$ that satisfies (LP$_0$) is a matter of ongoing research.

6.1 Consequences of state-dependent (GP)

In [2] it’s argued that $I_{\text{min}}$ upperbounds the ideal $I_\cap$. Inspired by $I_{\text{min}}$ assuming state-dependent (SR) and (M$_0$) to achieve a tighter upperbound on $I_\cap$, we assume state-dependent (GP) to achieve a tighter lowerbound on $I_\cap$ for $n = 2$. Our bound, denoted $I_{\text{sum}}$ for “sum minus pair”, is defined as,

$$I_{\text{sum}}(X_1, X_2 : Y) = \sum_{y \in Y} \Pr(y) \max \left[0, I(X_1 : y) + I(X_2 : y) - I(X_1 \cap X_2 : y)\right], \quad (10)$$

where $I(\bullet : y)$ is the same Kullback-Liebler divergence from eq. (1).

For example SUBTLE, the target $Y \equiv X_1 \cap X_2$, therefore per (Id), $I_\cap(X_1, X_2 : Y) = I(X_1 : X_2) = 0.252$ bits. However, given state-dependent (GP), applying $I_{\text{sum}}$ yields $I_\cap(X_1, X_2 : Y) \geq 0.390$. Therefore, (Id) and state-dependent (GP) are incompatible. Secondly, given state-dependent (GP), example SUBTLE additionally illustrates a conjecture from [4] that the intersection information two predictors have about a target can exceed the mutual information between them, i.e., $I_\cap(X_1, X_2 : Y) \neq I(X_1 : X_2)$.

7 Conclusion and Path Forward

We’ve made incremental progress on several fronts towards the ideal Shannon $I_\cap$.

**Desired Properties.** We have tightened, expanded, and pruned the desired properties for $I_\cap$. Particularly,

- (LB) is a non-contentious yet tighter lower-bound on $I_\cap$ than (GP).
- Motivated by the natural equality $I_\cap(X_1, \ldots, X_n : Y) = I_\cap(X_1, \ldots, X_n, Y : Y)$, we introduce (M$_1$) as a desired property.
- What was before an implicit assumption, we introduce (Eq) to better ground one’s thinking.

---

$^{10}$ $I_\cap$ and $I_\cup$ are duals related by the inclusion-exclusion principle. For arbitrary $n$, this is $I_\cup(X_1, \ldots, X_n : Y) = \sum_{S \subseteq \{X_1, \ldots, X_n\}} (-1)^{|S|+1} I_\cap(S_1, \ldots, S_{|S|} : Y)$. 

---

10
(a) $\text{Pr}(x_1, x_2, y)$

<table>
<thead>
<tr>
<th>$X_1 X_2$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td>00</td>
</tr>
<tr>
<td>1 1</td>
<td>11</td>
</tr>
<tr>
<td>0 1</td>
<td>01</td>
</tr>
</tbody>
</table>

$I(\text{X}_1, \text{X}_2 : Y) = 1.585$
$I(\text{X}_1 : Y) = 0.918$
$I(\text{X}_2 : Y) = 0.918$
$I(\text{X}_1 : \text{X}_2) = 0.252$
$I_{\min}(\text{X}_1, \text{X}_2 : Y) = 0.585$
$I_{\lambda}(\text{X}_1, \text{X}_2 : Y) = 0.0$
$I_{\text{smp}}(\text{X}_1, \text{X}_2 : Y) = 0.390$

(b) Circuit diagram

(c) $I_{\min}$
(d) $I_{\lambda}$ and $I_{\lambda}^0$
(e) $I_{\text{smp}}$

Figure 4: Example Subtle. In this example both $I_{\lambda}$ and $I_{\lambda}^0$ calculate a synergy of $-0.252$ bits of synergy. What kind of redundancy must be captured for a nonnegative decomposition for this example?
• A separate chain-rule property is superfluous. Any desirable properties of conditional $I_1$ are simply consequences of $(GP)$ and $(TM)$.

**A new measure.** Based on the Gács-Körner common random variable, we introduced a new Shannon $I_\lambda$ measure. Our measure, $I_\lambda$, is theoretically principled and the first to satisfy $(TM)$.

**How to improve.** We identified where $I_\lambda$ fails; it does not detect “imperfect” correlations between $X_1$ and $X_2$. One next step is to develop a less stringent $I_\lambda$ measure that satisfies $(LP_0)$ for simple nondeterministic examples like IMPERFECTRDN while still satisfying $(TM)$.

To our surprise, example SUBTLE shows that $I_\lambda^0$ does not satisfy $(LP_0)$! This suggests that $I_\lambda^0$ is too strict—what kind of zero-error informational overlap is $I_\lambda^0$ not capturing? A separate next step is to formalize what exactly is required for a zero-error $I_\lambda$ to satisfy $(LP_0)$.

Finally, we showed that state-dependent $(GP)$, a seemingly reasonable property, is incompatible with $(Id)$ and moreover entails that $I_\lambda(X_1,X_2;Y)$ can exceed $I(X_1:X_2)$.

**Acknowledgements.** VG thanks Tracey Ho and EKPC thanks Hua Li for valuable discussions. While intrepidly pulling back the veil of ignorance, VG was funded by a DOE CSGF fellowship. EKPC was funded by the CSU Information Science and Technology Center (ISTeC). RGJ and JPC were funded by ARO grant W911NF-12-1-0234, and CJE by a grant from the John Templeton Foundation.

**References**


**Appendix**

**Appendix A Algorithm for Computing Common Random Variable**

Given $n$ random variables $X_1, \ldots, X_n$, the common random variable $X_1 \wedge \cdots \wedge X_n$ is computed by steps 1–3 in Appendix B.
B Algorithm for Computing $I_{\lambda}$

1. For each $X_i$, for $i = 1, \ldots, n$, take its states $x_i$ and place them as nodes on a graph. At the end of this process there will be $\sum_{i=1}^n |X_i|$ nodes on the graph.

2. For each pair of RVs $X_i, X_j$ ($i \neq j$), draw an undirected edge connecting nodes $x_i$ and $x_j$ if $Pr(x_i, x_j) > 0$. At the end of this process the undirected graph will consist of $k$ connected components $1 \leq k \leq \min_i |X_i|$. Denote these $k$ disjoint components as $c_1, \ldots, c_k$.

3. Each connected component of the graph constitutes a distinct state of the common random variable $Q$, i.e., $|Q| = k$. Denote the states of the common random variable $Q$ by $q_1, \ldots, q_k$.

4. Construct the joint probability distribution $Pr(Q, Y)$ as follows. For every state $(q_i, y) \in Q \times Y$, the joint probability is created by summing over the entries of $Pr(x_1, \ldots, x_n, y)$ in component $i$. More precisely,

$$Pr(Q = q_i, Y = y) = \sum_{x_1, \ldots, x_n} Pr(x_1, \ldots, x_n, y) \quad \text{if } \{x_1, \ldots, x_n\} \subseteq c_i.$$

5. Using $Pr(Q, Y)$, compute $I_{\lambda}(X_1, \ldots, X_n : Y)$ simply by computing the Shannon mutual information between $Q$ and $Y$, i.e., $I(Q : Y) = D_{KL}[Pr(Q, Y) \| Pr(Q) Pr(Y)]$.

C Lemmas and Proofs

C.1 Lemmas on Desired Properties

Lemma 2. If (LB) holds, then $I_{\gamma}(X_1, \ldots, X_n : Y) \geq I(X_1 \wedge \cdots \wedge X_n : Y)$.

Proof. Assume that (LB) holds. By definition, $X_1 \wedge \cdots \wedge X_n \preceq X_i$ for $i = 1, \ldots, n$. So, by (LB), we immediately conclude that $I_{\gamma}(X_1, \ldots, X_n : Y) \geq I(X_1 \wedge \cdots \wedge X_n : Y)$, which is the desired result.

For the converse, we need the following assumption:

(IM) If $X_1 \succeq X_2$, then $I(X_1 : Y) \leq I(X_2 : Y)$.

Lemma 3. Suppose that (IM) holds, and that $I_{\gamma}(X_1, \ldots, X_n : Y) \geq I(X_1 \wedge \cdots \wedge X_n : Y)$. Then (LB) holds.

Proof. Assume that $I_{\gamma}(X_1, \ldots, X_n : Y) \geq I(X_1 \wedge \cdots \wedge X_n : Y)$. Let $Q \preceq X_i$ for $i = 1, \ldots, n$. Because $X_1 \wedge \cdots \wedge X_n$ is the largest (informationally richest) random variable that is informationally poorer than $X_i$ for $i = 1, \ldots, n$, it follows that $Q \not\preceq X_1 \wedge \cdots \wedge X_n$. Therefore, by (IM), $I(X_1 \wedge \cdots \wedge X_n : Y) \geq I(Q : Y)$. Hence, $I_{\gamma}(X_1, \ldots, X_n : Y) \geq I(Q : Y)$ also, which completes the proof.

Remark: Assumption (IM) is satisfied by zero-error information and Shannon mutual information.

Lemma 4. Given $I_{\gamma}, X_1, \ldots, X_n, Y, \text{ and } Z$, consider the conditional intersection information

$$I_{\gamma}(X_1, \ldots, X_n : Z | Y) = I_{\gamma}(X_1, \ldots, X_n : Y \uplus Z) - I_{\gamma}(X_1, \ldots, X_n : Y).$$

Suppose that (GP), (Eq), and (TM) hold. Then, the following properties hold:

- $I_{\gamma}(X_1, \ldots, X_n : Z | Y) \geq 0$.

- $I_{\gamma}(X_1, \ldots, X_n : Z | Y) = I_{\gamma}(X_1, \ldots, X_n : Z)$ if $Y$ is a constant.
Thus, I hold for arbitrary arbitrary. However, this inequality does not hold in general. To see this, suppose that it does

as desired.

C.2 Properties of \( I^0 \)

**Lemma 5.** The measure of intersection information \( I^0_\lambda(X_1, \ldots, X_n : Y) \) satisfies (GP), (Eq), (TM), (M0), and (S0), but not (LP0).

**Proof.** (GP): The inequality \( I^0_\lambda(X_1, \ldots, X_n : Y) \geq 0 \) follows immediately from the identity \( I^0_\lambda(X_1, \ldots, X_n : Y) = H(X_1 \cdots \cup X_n, Y) \) and the nonnegativity of H(\cdot). Next, if Y is a constant, then by the generalized absorption law, \( X_1 \cdots \cup X_n, Y \equiv Y \). Thus, by the invariance of H(\cdot) (Lemma 1(a)), \( H(X_1 \cdots \cup X_n, Y) = H(Y) = 0 \).

(Eq): Consider \( X_1 \cdots \cup X_n, Y \). The equivalence class (with respect to \( \equiv \)) in which

this random variable resides is closed under substitution of \( X_i \) (for \( i = 1, \ldots, n \)) or Y by an informationally equivalent random variable. Hence, because \( I^0_\lambda(X_1, \ldots, X_n : Y) = H(X_1 \cdots \cup X_n, Y) \) and H(\cdot) is invariant over the equivalence class of random variables that are informationally equivalent to \( X_1 \cdots \cup X_n, Y \) (by Lemma 1(a)), the desired result holds.

(TM): Suppose that \( Y \preceq Z \). Then, \( X_1 \cdots \cup X_n, Y \preceq X_1 \cdots \cup X_n, Z \). Then, we have

\[
I^0_\lambda(X_1, \ldots, X_n : Y) = H(X_1 \cdots \cup X_n, Y) \\
\leq H(X_1 \cdots \cup X_n, Z) \quad \text{by monotonicity of } H(\cdot) \text{ (Lemma 1(b))} \\
= I^0_\lambda(X_1, \ldots, X_n : Z),
\]

as desired.

(M0): By the generalized absorption law, \( X_1 \cdots \cup X_n, W \preceq Y \preceq X_1 \cdots \cup X_n, Y \). Hence,

\[
I^0_\lambda(X_1, \ldots, X_n, W : Y) = H(X_1 \cdots \cup X_n, W, Y) \\
\leq H(X_1 \cdots \cup X_n, Y) \quad \text{by monotonicity of } H(\cdot) \text{ (Lemma 1(b))} \\
= I^0_\lambda(X_1, \ldots, X_n : Y),
\]

as desired.

Next, suppose that there exists \( Z \in \{X_1, \ldots, X_n\} \) such that \( Z \preceq W \). Then, by the generalized absorption law, \( X_1 \cdots \cup X_n, W, Y \equiv X_1 \cdots \cup X_n, Y \). Hence,

\[
I^0_\lambda(X_1, \ldots, X_n, W : Y) = H(X_1 \cdots \cup X_n, W, Y) \\
= H(X_1 \cdots \cup X_n, Y) \quad \text{by invariance of } H(\cdot) \text{ (Lemma 1(a))} \\
= I^0_\lambda(X_1, \ldots, X_n : Y),
\]

as desired.

(S0): By the commutativity law, \( X_1 \cdots \cup X_n, Y \) is invariant (with respect to \( \equiv \)) under reordering of \( X_1, \ldots, X_n \). Hence, the desired result follows immediately from the identity \( I^0_\lambda(X_1, \ldots, X_n : Y) = H(X_1 \cdots \cup X_n, Y) \) and the invariance of H(\cdot) (Lemma 1(a)).

(LP0): For \( I^0_\lambda, (LP0) \) relative to zero-error information can be written as

\[
H(X_1 \cup X_2, Y) \geq H(X_1, Y) + H(X_2, Y) - H((X_1 \cap X_2), Y).
\]

However, this inequality does not hold in general. To see this, suppose that it does hold for arbitrary \( X_1, X_2, \) and \( Y \). Note that \( (X_1 \cap X_2) \cup Y \preceq Y \), which implies that

\[
H((X_1 \cap X_2) \cup Y) \leq H(Y) \quad \text{(by monotonicity of } H(\cdot)\text{). Hence, the inequality (11) implies that}

\[
H(X_1 \cup X_2, Y) \geq H(X_1, Y) + H(X_2, Y) - H(Y).
\]
Rewriting this, we get
\[ H(X_1 \land Y) + H(Y \land X_2) \leq H(X_1 \land Y \land X_2) + H(Y). \]

But this is the supermodularity law for common information, which is known to be false in general; see [6], Section 5.4.

Lemma 6. With respect to zero-error information, the measure of intersection information \( I^0(X_1, \ldots, X_n : Y) \) satisfies (LB), (SR), and (Id).

Proof. (LB): Suppose that \( Q \preceq X_i \) for \( i = 1, \ldots, n \). Because \( X_1 \land \cdots \land X_n \) is the largest (informationally richest) random variable that is informationally poorer than \( X_i \) for \( i = 1, \ldots, n \), it follows that \( Q \preceq X_1 \land \cdots \land X_n \). This implies that \( I^0(X_1 \land \cdots \land X_n \land Y) \geq I^0(Q : Y) \). Therefore,
\[
I^0(X_1, \ldots, X_n : Y) = H(X_1 \land \cdots \land X_n \land Y) \\
\geq H(Q \land Y) \quad \text{by monotonicity of } H(\cdot) \quad \text{(Lemma 1(b))} \\
= I^0(Q : Y),
\]
as desired.

(SR): We have \( I^0(X_1 : Y) = H(X_1 \land Y) = I^0(X_1 : Y) \).

(Id): By the associative and absorption laws, we have \( X \land Y \land (X \lor Y) \cong X \land Y \). Thus,
\[
I^0(X, Y : X \lor Y) = H(X \land Y \land (X \lor Y)) \\
= H(X \land Y) \quad \text{by invariance of } H(\cdot) \quad \text{(Lemma 1(a))} \\
= I^0(X : Y),
\]
as desired.

Lemma 7. The measure of intersection information \( I^0(X_1, \ldots, X_n : Y) \) satisfies (M) and (S), but not (LP).

Proof. (M): The desired inequality is identical to (M), so it remains to prove the sufficient condition for equality. Suppose that there exists \( Z \in \{X_1, \ldots, X_n, Y\} \) such that \( Z \preceq W \). Then, by the generalized absorption law, \( X_1 \land \cdots \land X_n \land W \land Y \cong X_1 \land \cdots \land X_n \land Z \). Hence,
\[
I^0(X_1, \ldots, X_n, W : Y) = H(X_1 \land \cdots \land X_n \land W \land Y) \\
= H(X_1 \land \cdots \land X_n \land Z) \quad \text{by invariance of } H(\cdot) \quad \text{(Lemma 1(a))} \\
= I^0(X_1, \ldots, X_n : Z),
\]
as desired.

(S): By the commutativity law, \( X_1 \land \cdots \land X_n \land Y \) is invariant (with respect to \( \cong \)) under reordering of \( X_1, \ldots, X_n, Y \). Hence, the desired result follows immediately from the identity \( I^0(X_1, \ldots, X_n : Y) = H(X_1 \land \cdots \land X_n \land Y) \) and the invariance of \( H(\cdot) \) (Lemma 1(a)).

(LP): This follows from not satisfying (LP).

C.3 Properties of \( I_\lambda \)

Lemma 8. The measure of intersection information \( I_\lambda(X_1, \ldots, X_n : Y) \) satisfies (GP), (Eq), (TM), (M), and (S), but not (LP).
Proof. (GP): The inequality $I_\lambda(X_1, \ldots, X_n : Y) \geq 0$ follows immediately from the identity $I_\lambda(X_1, \ldots, X_n : Y) = I(X_1 \wedge \cdots \wedge X_n : Y)$ and the nonnegativity of mutual information. Next, suppose that $Y$ is a constant. Then $H(Y) = 0$. Moreover, $Y \lessdot X_1 \wedge \cdots \wedge X_n$ by definition of $\lessdot$. Thus, by Lemma 1(c), $H(Y|X_1 \wedge \cdots \wedge X_n) = 0$, and

$$I_\lambda(X_1, \ldots, X_n : Y) = I(X_1 \wedge \cdots \wedge X_n : Y) = I(Y : X_1 \wedge \cdots \wedge X_n) = H(Y) - H(Y|X_1 \wedge \cdots \wedge X_n) = 0.$$ (Eq)

Consider $X_1 \wedge \cdots \wedge X_n \wedge Y$. The equivalence class (with respect to $\equiv$) in which this random variable resides is closed under substitution of $X_i$ (for $i = 1, \ldots, n$) or $Y$ by an informationally equivalent random variable. Hence, because

$$I_\lambda(X_1, \ldots, X_n : Y) = H(Y) - H(Y|X_1 \wedge \cdots \wedge X_n) = H(X_1 \wedge \cdots \wedge X_n) - H(X_1 \wedge \cdots \wedge X_n|Y),$$

by Lemma 1(a), the desired result holds.

(TM): Suppose that $Y \lessdot Z$. For simplicity, let $Q = X_1 \wedge \cdots \wedge X_n$. Then,

$$I_\lambda(X_1, \ldots, X_n : Y) = H(Q) - H(Q|Y) \leq H(Q) - H(Q|Z) \quad \text{by Lemma 1(b)}$$

$$= I_\lambda(X_1, \ldots, X_n : Z),$$

as desired.

(M0): By definition of $\wedge$, we have $X_1 \wedge \cdots \wedge X_n \wedge W \lessdot X_1 \wedge \cdots \wedge X_n$. Hence,

$$I_\lambda(X_1, \ldots, X_n, W : Y) = H(X_1 \wedge \cdots \wedge X_n \wedge W) - H(X_1 \wedge \cdots \wedge X_n \wedge W|Y) \leq H(X_1 \wedge \cdots \wedge X_n) - H(X_1 \wedge \cdots \wedge X_n|Y) \quad \text{by Lemma 1(b)}$$

$$= I_\lambda(X_1, \ldots, X_n : Y),$$

as desired.

Next, suppose that there exists $Z \in \{X_1, \ldots, X_n\}$ such that $Z \lessdot W$. Then, by the algebraic laws of $\wedge$, we have $X_1 \wedge \cdots \wedge X_n \wedge W \equiv X_1 \wedge \cdots \wedge X_n$. Hence,

$$I_\lambda(X_1, \ldots, X_n, W : Y) = H(X_1 \wedge \cdots \wedge X_n \wedge W) - H(X_1 \wedge \cdots \wedge X_n \wedge W|Y) = H(X_1 \wedge \cdots \wedge X_n) - H(X_1 \wedge \cdots \wedge X_n|Y) \quad \text{by Lemma 1(a)}$$

$$= I_\lambda(X_1, \ldots, X_n : Y),$$

as desired.

(S0): By the commutativity law, $X_1 \wedge \cdots \wedge X_n$ is invariant (with respect to $\equiv$) under reordering of $X_1, \ldots, X_n$. Hence, the desired result follows immediately from the identity $I_\lambda(X_1, \ldots, X_n : Y) = H(X_1 \wedge \cdots \wedge X_n) - H(X_1 \wedge \cdots \wedge X_n|Y)$ and Lemma 1(a).

(LP0): A counterexample is provided by ImperfectRdn (Figure 3).

Lemma 9. With respect to mutual information, the measure of intersection information $I_\lambda(X_1, \ldots, X_n : Y)$ satisfies (LB) and (SR), but not (Id).

Proof. (LB): Suppose that $Q \lessdot X_i$ for $i = 1, \ldots, n$. Because $X_1 \wedge \cdots \wedge X_n$ is the largest (informationally richest) random variable that is informationally poorer than $X_i$ for $i = 1, \ldots, n$, it follows that $Q \lessdot X_1 \wedge \cdots \wedge X_n$. Therefore,

$$I_\lambda(X_1, \ldots, X_n : Y) = H(X_1 \wedge \cdots \wedge X_n) - H(X_1 \wedge \cdots \wedge X_n|Y) \geq H(Q) - H(Q|Y) \quad \text{by Lemma 1(b)}$$

$$= I(Q : Y),$$

as desired.
as desired.

(SR): By definition, \( I_\lambda(X_1 : Y) = I(X_1 : Y) \).

(Id): We have \( X \times Y \preceq X \times Y \) by definition of \( \times \) and \( \preceq \). Thus,

\[
I_\lambda(X, Y : X \times Y) = I(X \times Y : X \times Y)
= H(X \times Y) - H(X \times Y | X \times Y)
= H(X \times Y) \quad \text{by Lemma 1(a)}
= I^0(X : Y)
\neq I(X : Y).
\]

\[\square\]

**Lemma 10.** The measure of intersection information \( I_\lambda(X_1, \ldots, X_n : Y) \) does not satisfy (M1), (S1), and (LP1).

**Proof.** (M1): A counterexample is provided in IMPERFECTRDN (Figure 3), where \( I_\lambda(X_1 : Y) = 0.99 \) bits, yet \( I_\lambda(X_1, Y : Y) = 0 \) bits.

(S1): A counterexample. We show \( I_\lambda(X, X : Y) \neq I_\lambda(X, Y : X) \).

\[
I_\lambda(X, X : Y) - I_\lambda(X, Y : X)
= I(X : Y) - I(X, Y : X)
= I(X : Y) - I(X \times Y : X)
= I(X : Y) - H(X \times Y) - H(X \times Y | X)
= I(X : Y) - H(X \times Y)
\neq 0.
\]

(LP1): This follows from not satisfying (LP0).

\[\square\]

**D Miscellaneous Results**

**Simplification of \( I_\lambda^0 \)**

**Lemma 11.** We have \( I_\lambda^0(X_1, \ldots, X_n : Y) = H(X_1 \times \cdots \times X_n \times Y) \).

**Proof.** By definition,

\[
I_\lambda^0(X_1, \ldots, X_n : Y) = \max_{\Pr(Q | Y)} I^0(Q : Y)
\quad \text{subject to } Q \preceq X_i \forall i \in \{1, \ldots, n\}
= \max_{\Pr(Q | Y)} H(Q \times Y)
\quad \text{subject to } Q \preceq X_i \forall i \in \{1, \ldots, n\}
\]

Let \( Q \) be an arbitrary random variable satisfying the constraint \( Q \preceq X_i \) for \( i = 1, \ldots, n \). Because \( X_1 \times \cdots \times X_n \) is the largest random variable (in the sense of the partial order \( \preceq \)) that is informationally poorer than \( X_i \) for \( i = 1, \ldots, n \), we have \( Q \preceq X_1 \times \cdots \times X_n \times Y \). By the property of \( \times \) pointed out before, we also have \( Q \times Y \preceq X_1 \times \cdots \times X_n \times Y \). By Lemma 1(b), this implies that \( H(Q \times Y) \leq H(X_1 \times \cdots \times X_n \times Y) \). Therefore, \( I_\lambda^0(X_1, \ldots, X_n : Y) = H(X_1 \times \cdots \times X_n \times Y) \).

\[\square\]

**Simplification of \( I_\lambda \)**

**Lemma 12.** We have \( I_\lambda(X_1, \ldots, X_n : Y) = I(X_1 \times \cdots \times X_n : Y) \).
Proof. By definition,

$$I_\lambda(X_1,\ldots,X_n;Y) \equiv \max_{\Pr(Q|Y)} I(Q;Y)$$

subject to $Q \leq X_i \forall i \in \{1,\ldots,n\}$

$$= H(Y) - \min_{\Pr(Q)} H(Y|Q)$$

subject to $Q \leq X_i \forall i \in \{1,\ldots,n\}$

Let $Q$ be an arbitrary random variable satisfying the constraint $Q \leq X_i$ for $i = 1,\ldots,n$. Because $X_1 \land \cdots \land X_n$ is the largest random variable (in the sense of the partial order $\leq$) that is informationally poorer than $X_i$ for $i = 1,\ldots,n$, we have $Q \leq X_1 \land \cdots \land X_n$. By Lemma 1(b), this implies that $H(Y|Q) \geq H(Y|X_1 \land \cdots \land X_n \land Y)$. Therefore, $I_\lambda(X_1,\ldots,X_n;Y) = I(X_1 \land \cdots \land X_n;Y).

Proof that $I_\lambda(X_1,\ldots,X_n;Y) \leq I_{\text{min}}(X_1,\ldots,X_n;Y)$

Lemma 13. We have $I_\lambda(X_1,\ldots,X_n;Y) \leq I_{\text{min}}(X_1,\ldots,X_n;Y)$

Proof. Starting from the definitions,

$$I_\lambda(X_1,\ldots,X_n;Y) \equiv I(X_1 \land \cdots \land X_n;Y)$$

$$= \sum_y \Pr(y) I(X_1 \land \cdots \land X_n:y)$$

$$I_{\text{min}}(X_1,\ldots,X_n;Y) \equiv \sum_y \Pr(y) \min_i I(X_i:y).$$

For a particular state $y$, without loss of generality we define the minimizing predictor $X_m$ by $X_m \equiv \arg\min_y I(X_i;y)$ and the common random variable $Q \equiv X_1 \land \cdots \land X_n$. It then remains to show that $I(Q:y) \leq I(X_m:y)$.

By definition of $\land$, we have $Q \leq X_m$. Hence,

$$I(X_m:y) = H(X_m) - H(X_m|Y = y)$$

$$\geq H(Q) - H(Q|Y = y) \quad \text{by Lemma 1(b)}$$

$$= I(Q:y).$$

State-dependent zero-error information

We define the state-dependent zero-error information, $\Gamma^0(X:Y = y)$ as,

$$\Gamma^0(X:Y = y) \equiv \log \frac{1}{\Pr(Q = q)},$$

where the random variable $Q \equiv X \land Y$ and $\Pr(Q = q)$ is the probability of the connected component containing state $y \in Y$. This entails that $\Pr(y) \leq \Pr(q) \leq 1$. Similar to the state-dependent information, $\mathbb{E}_Y \Gamma^0(X:y) = \Gamma^0(X:Y)$, where $\mathbb{E}_Y$ is the expectation value over $Y$.

Proof. We define two functions $f$ and $g$:

- $f : y \to q$ s.t. $\Pr(q|y) = 1$ where $q \in Q$ and $y \in Y$.
- $g : q \to \{y_1,\ldots,y_k\}$ s.t. $\Pr(q|y_k) = 1$ where $q \in Q$ and $y \in Y$. 

18
Now we have,

\[ \mathbb{E}_Y I^0(X : Y) = \sum_{y \in Y} \Pr(y) \log \frac{1}{\Pr(f(y))} \cdot \]

Since each \(y\) is associated with exactly one \(q\), we can reindex the \(\sum_{y \in Y}\). We then simplify to achieve the result.

\[
\sum_{y \in Y} \Pr(y) \log \frac{1}{\Pr(f(y))} = \sum_{q \in Q} \sum_{y \in g(q)} \Pr(y) \log \frac{1}{\Pr(f(y))}
\]

\[
= \sum_{q \in Q} \sum_{y \in g(q)} \Pr(y) \log \frac{1}{\Pr(q)} = \sum_{q \in Q} \log \frac{1}{\Pr(q)} \sum_{y \in g(q)} \Pr(y)
\]

\[
= \sum_{q \in Q} \log \frac{1}{\Pr(q)} \Pr(q) = \sum_{q \in Q} \Pr(q) \log \frac{1}{\Pr(q)}
\]

\[
= H(Q) = I^0(X : Y).
\]
Figure 5: Example RDN. In this example $I_{\text{min}}$ and $I_{\lambda}$ reach the same answer yet diverge drastically for example \textsc{ImperfectRdn}.