Pooling probability distributions and partial information decomposition

S. J. van Enk
Department of Physics, University of Oregon, Eugene, Oregon 97403, USA

(Received 31 January 2023; accepted 9 May 2023; published 25 May 2023)

Notwithstanding various attempts to construct a partial information decomposition (PID) for multiple variables by defining synergistic, redundant, and unique information, there is no consensus on how one ought to precisely define either of these quantities. One aim here is to illustrate how that ambiguity—or, more positively, freedom of choice—may arise. Using the basic idea that information equals the average reduction in uncertainty when going from an initial to a final probability distribution, synergistic information will likewise be defined as a difference between two entropies. One term is uncontroversial and characterizes "the whole" information that source variables carry jointly about a target variable $\mathcal{T}$. The other term then is meant to characterize the information carried by the "sum of its parts." Here we interpret that concept as needing a suitable probability distribution aggregated ("pooled") from multiple marginal distributions (the parts). Ambiguity arises in the definition of the optimum way to pool two (or more) probability distributions. Independent of the exact definition of optimum pooling, the concept of pooling leads to a lattice that differs from the often-used redundancy-based lattice. One can associate not just a number (an average entropy) with each node of the lattice, but (pooled) probability distributions. As an example, one simple and reasonable approach to pooling is presented, which naturally gives rise to the overlap between different probability distributions as being a crucial quantity that characterizes both synergistic and unique information.

DOI: 10.1103/PhysRevE.107.054133

I. INTRODUCTION

Since the seminal work by Williams and Beer [1] on the partial information decomposition (PID) and their proposed definitions for synergistic, redundant, and unique information, a lot of progress has been made to further clarify these notions. While the intuitive notions seem fairly clear at first sight, upon closer study there is some ambiguity left that has been surprisingly difficult to eliminate, as witnessed by the many valuable but different proposals for defining an explicit PID [2–10]. One point of the current work is to locate where that ambiguity may arise. The other is to question whether all terms in the decomposition correspond to information.

There are various reasons for desiring a PID. Reference [11] gives a nice example of two probability distributions over three variables that, although produced by clearly distinct mechanisms, cannot be distinguished by just using the standard mutual information between the different combinations of the three variables. But since those two distributions can be distinguished by using a PID, this demonstrates an immediate use of that decomposition as revealing different underlying mechanisms. In that role, a PID may help our understanding of complex networks [12,13], such as neural networks, either involving actual neurons or artificial ones [14–16].

Moreover, synergistic information may explain emergence [17,18] as implementing the adage that "the whole is more than the sum of its parts." Perhaps a similar idea may even explain or define consciousness [19]. For a perspective on the uses and prospects of PID, see Ref. [20], and for discussions of important ideas that are made use of in the following, see Refs. [3,5,21]. For earlier definitions of synergy and redundancy in the context of neural encoding outside the framework of PID, see Refs. [22,23].

Here we propose a new perspective. Just like Ref. [24] went back to basics about part-whole relationships in order to get a fundamental idea of synergistic information, here we return to basics about information. Within Shannon's theory [25,26], information is defined as the average reduction in uncertainty upon changing from an initial probability distribution to a final distribution. Information is then always of the generic form

$$\text{Information} = H_{\text{init}} - H_{\text{final}} \tag{1}$$

as a difference between two (averaged) entropies. Moreover, information is always *about* something, and in our case the something will be represented by a target variable $\mathcal{T}$. The uncertainty about $\mathcal{T}$ taking on a particular value t is quantified as $\log(1/\mathrm{P}(t))$, with the logarithm taken in base 2, and with $\mathrm{P}(t)$ the probability of finding the value t. This uncertainty is averaged over all possible values $\mathcal{T}$ can take and over other initial and final variables, respectively, thus yielding expressions for the entropies H_{init} and H_{final} in Eq. (1). Typically, the relation between final and initial probability distributions is such that information is nonnegative.

Using this basic idea, in order to define synergistic information, we will need, besides the joint distribution, a single distribution aggregated from multiple marginal distributions. To define such a distribution it seems easiest to use the language and techniques of "pooling distributions" for combining different expert opinions into one opinion (by compromise if not by consensus) [27–30]. This language, although not necessary, is convenient for setting up notation, as explained in Secs. II A and II B. In Sec. II C we define synergistic information in terms of a generic pooled probability distribution. We provide a side-by-side comparison of our PID with one concrete measure, taken from Ref. [3], as both the analogies and the differences are clear and illuminating. Another comparison can be made by setting up an alternative PID by defining redundant information (as a difference between two averaged entropies) first.

The idea of pooling will help us define (in Sec. II D) a *lattice* [31] underlying the structure of the PID for multiple variables, which differs from the redundancy-based lattice used in the original Williams-Beer work [1].

Concrete ways of pooling probability distributions are discussed in Sec. III. For one convenient and popular way of pooling, synergistic information and unique information acquire natural interpretations, as shown in Sec. III B. Examples illustrating the consequences for a PID based on pooling are given in Sec. IV.

II. SETUP AND NOTATION

A. Experts and their opinions

We may use the following scenario to set terminology and notation. Capital letters indicate variables, lower-case letters indicate their possible values. Entropies and information are expressed in units of bits.

Suppose we are interested in a variable $\mathcal{T}$. We hire several experts (or agents), Alice, Bob, Charlie, ...to help us predict which of possible (discrete) values $t_1, t_2, \dots$ $\mathcal{T}$ may take on under specific conditions. More precisely, each expert X is expected to report to us a probability distribution $\mathrm{P}_X(\mathcal{T})$.

Each expert has their own laboratory and each performs their own measurements of a single variable, which is indicated by the agent's initial $A, B, C, \dots$. Each expert measures how their variable correlates with $\mathcal{T}$. That is, each expert estimates a probability distribution $P_X(\mathcal{T}) := P(\mathcal{T}|X)$ for $X = A, B, C, \dots$ [32]. Expert X's uncertainty regarding $\mathcal{T}$ is quantified by the Shannon entropy [25] of the conditional distribution

$$H_{\{X\}} = H(\mathcal{T}|X) = \sum_x \sum_t \mathrm{P}(x,t) \log(1/\mathrm{P}(t|x)). \qquad (2)$$

The subscript uses set notation, this being convenient for the general definition of the PID. In this case, we have one set containing one expert, X. We may make this definition of uncertainty operational by imagining that we impose a fine on expert X equal to

$$\mathrm{Fine}_X(t) = \log(1/\mathrm{P}_X(t)) \qquad (3)$$

if the value t actually occurs. (The concept behind this operational definition goes by the name of "scoring rules" and one important feature of this specific rule is that the expert is forced to report their best (or "true") probability distribution if they want to minimize their fine [33].)

We assume here consistency among all experts, in that the unconditioned distribution for $\mathcal{T}$ is identical for all agents. In other words, we assume here that there is a joint distribution $\mathrm{P}(\mathcal{T}, A, B, C \dots)$ from which all other distributions can be obtained by marginalizing over some or all of the variables $A, B, C \dots$ [34]. In particular, we may define

$$\mathrm{P}_\varnothing(t) = \sum_a \mathrm{P}_A(t|a) = \sum_b \mathrm{P}_B(t|b) = \dots \qquad (4)$$

with the empty set in the subscript indicating we need no expert opinions for this probability distribution. The uncertainty we have before learning from any expert is likewise denoted by

$$H_\varnothing = H(\mathcal{T}). \qquad (5)$$

To quantify how much information each expert provides to us individually we use the standard measure for mutual information

$$I_{\{X\}} = H_\varnothing - H_{\{X\}}, \qquad (6)$$

in line with the basic definitional form (1) for information. Operationally, this equals the amount of money the expert saves by reducing their fine, by reporting their probability distribution P_X rather than $\mathrm{P}_\varnothing$.

Our experts may collaborate, as follows. Each expert, say, Alice, Bob, and Charlie, still just measures their own variable, but by communicating with each other and synchronizing their measurements, they may find the joint distribution $\mathrm{P}(\mathcal{T}, A, B, C)$. They can thereby jointly report $P(\mathcal{T}|ABC)$. The mutual information between the joint variables and $\mathcal{T}$ is then denoted by $I_{\{A,B,C\}}$:

$$I_{\{A,B,C\}} = H_\varnothing - H_{\{A,B,C\}}. \qquad (7)$$

Operationally, this equals the amount of money the three experts save themselves by collaborating. The second term involves a single probability distribution for each set of given values a, b, c of the variables A, B, C.

B. Pooling experts' opinions

There is another way of producing a single distribution for $\mathcal{T}$ by pooling the individual experts' "opinions" $\mathrm{P}(\mathcal{T}|x)$ for $x = a, b, c$. One very simple (albeit inadequate) way to do this, would be to use the average distribution

$$\mathrm{P}_{\{A\},\{B\},\{C\}}(\mathcal{T}) = \tfrac{1}{3}[\mathrm{P}_A(\mathcal{T}) + \mathrm{P}_B(\mathcal{T}) + \mathrm{P}_C(\mathcal{T})], \qquad (8)$$

but, clearly, there are many more. The operational idea is that we ask the three experts to report to us a single distribution, even if they never collaborated. The rule of combining different distributions $\mathrm{P}(\mathcal{T}|X)$ should be symmetric between all experts, in order to conform to standard axioms required for a PID. Note that the subscript now contains three sets, containing one expert each, reflecting that the pooled distribution makes use only of single-expert opinions $\mathrm{P}_X(\mathcal{T})$. Once the idea of pooling is in place we can easily extend it to different types of combinations of experts. For example, $\mathrm{P}_{\{A,B\},\{C\}}$ would be constructed out of the joint distribution produced by Alice and Bob collaborating and Charlie's single-expert distribution P_C.

To make a distinction between the different types of combinations of experts it may be useful to talk about collections of sets of experts. For example, $\{A, B\}$, $\{C\}$ denotes a collection of two sets, one set containing two experts, the other containing one expert.

C. Synergistic, unique, and redundant information

1. Definitions

Now let us first focus on just a pair of experts, Alice and Bob, and in what way their pooled distribution $\mathrm{P}_{\{A\},\{B\}}(\mathcal{T})$ determines the PID. We first display the usual PID (sticking with standard notation for the moment) for how two variables carry information in different ways about $\mathcal{T}$:

$$\begin{aligned} I_{\{A,B\}} &= I_{\mathrm{unq}\,A\backslash B} + I_{\mathrm{unq}\,B\backslash A} + I_{\mathrm{red}\,A\&B} + I_{\mathrm{syn}\,A\&B},\\ I_{\{A\}} &= I_{\mathrm{unq}\,A\backslash B} + I_{\mathrm{red}\,A\&B},\\ I_{\{B\}} &= I_{\mathrm{unq}\,B\backslash A} + I_{\mathrm{red}\,A\&B}. \end{aligned} \tag{9}$$

In this notation, $I_{\mathrm{unq}\,A\backslash B}$ indicates information unique to Alice w.r.t. Bob, and $I_{\mathrm{red}\,A\&B}$ stands for the information redundantly encoded in (i.e., shared by) both Alice's and Bob's probability distributions $P_A(\mathcal{T})$ and $P_B(\mathcal{T})$, respectively.

We define synergistic information as the difference between two entropies:

$$I_{\mathrm{syn}\,A\&B} = H_{\{A\},\{B\}} - H_{\{A,B\}}, \tag{10}$$

as it captures how the whole (Alice and Bob collaborating and producing a joint distribution) is more than the sum of its parts (Alice and Bob pooling their individual distributions). But this then determines unique information as

$$\begin{aligned} I_{\mathrm{unq}\,A\backslash B} &= H_{\{B\}} - H_{\{A\},\{B\}},\\ I_{\mathrm{unq}\,B\backslash A} &= H_{\{A\}} - H_{\{A\},\{B\}}. \end{aligned} \tag{11}$$

This is again a difference between two entropies, and it gives the unique information that one expert possesses but the other does not. The operational definition is that it equals the money one expert can save for the other by pooling.

Finally, given a pooled distribution $\mathrm{P}_{\{A\},\{B\}}$ the redundant information is then given by

$$\Delta I_{\mathrm{red}\,A\&B} = H_{\{A\}\{B\}} + H_{\varnothing} - H_{\{A\}} - H_{\{B\}}. \tag{12}$$

This is not a difference between two entropies and it involves four (rather than two) different sorts of probability distributions over $\mathcal{T}$. It can, of course, be written as a difference between two information quantities. This is reflected in the use of the symbol $\Delta I_{\mathrm{red}\,A\&B}$. This notational device has been used before, for example in Refs. [22,23], to carefully distinguish different types of informational quantities and define synergy and redundancy in neural coding (see remark after the next equation).

We may also note that the above definition of redundant information is very similar to that of coinformation [35] {(with notation changed here by including the Δ symbol)

$$\Delta I_{\mathrm{co}\,A\&B} = H_{\{A,B\}} + H_{\varnothing} - H_{\{A\}} - H_{\{B\}}, \tag{13}$$

which features the joint distribution rather than the pooled distribution. It is well known that this quantity can take on both negative and positive values. This quantity, in fact, has been identified with synergy and it was denoted then as ΔI_{syn} in [23].

In order to see more completely which quantities are always differences between two entropies and which one may not, let us now also consider an alternative way (indicated by using primed symbols) of defining the PID by first defining redundant information as a difference between two (average) entropies. We use $H_{\varnothing}$ as the higher entropy and then need an entropy $H_{\mathrm{red}\,A\&B}$, symmetric between A and B, that derives from some distribution that contains only information that is common to both A and B, such that

$$I'_{\mathrm{red}\,A\&B} = H_{\varnothing} - H_{\mathrm{red}\,A\&B}. \tag{14}$$

We do not need to specify anything further in order to see that the unique information would then be given by

$$\begin{aligned} I'_{\mathrm{unq}\,A\backslash B} &= H_{\mathrm{red}\,A\&B} - H_{\{B\}},\\ I'_{\mathrm{unq}\,B\backslash A} &= H_{\mathrm{red}\,A\&B} - H_{\{A\}}. \end{aligned} \tag{15}$$

Just as before, unique information is then a difference between two entropies. On the other hand, the synergystic information would contain four terms and we would write

$$\Delta I_{\mathrm{syn}\,A\&B} = -H_{\{A,B\}} - H_{\mathrm{red}\,A\&B} + H_{\{A\}} + H_{\{B\}}. \tag{16}$$

In this case, then, synergy would be the odd one out, as being the only quantity in the PID that is not a difference between two entropies. Unique information is special, in that it is truly information in either of these two cases.

2. Relation to the BROJA measure

There are some similarities and contrasts between pooling and the way the BROJA measure—named for the authors of [3]—is defined. For the BROJA measure, one first considers all joint distributions $P_{\mathrm{poss}}(\mathcal{T}, A, B)$ that are consistent with the two marginal distributions $P(\mathcal{T}, A)$ and $P(\mathcal{T}, B)$. One then maximizes over all possible joint distributions the entropy of $P_{\mathrm{poss}}(\mathcal{T}|A, B)$. Denote that maximum by $\tilde{H}_{\{A\},\{B\}}$. The subscript here reminds us the maximum is determined by the two marginal distributions. Synergistic, unique, and redundant information are then given by

$$\begin{aligned} \tilde{I}_{\mathrm{syn}\,A\&B} &= \tilde{H}_{\{A\},\{B\}} - H_{\{A,B\}},\\ \tilde{I}_{\mathrm{unq}\,A\backslash B} &= H_{\{B\}} - \tilde{H}_{\{A\},\{B\}},\\ \tilde{I}_{\mathrm{unq}\,B\backslash A} &= H_{\{A\}} - \tilde{H}_{\{A\},\{B\}},\\ \Delta\tilde{I}_{\mathrm{red}\,A\&B} &= \tilde{H}_{\{A\}\{B\}} + H_{\varnothing} - H_{\{A\}} - H_{\{B\}}. \end{aligned} \tag{17}$$

Both the analogy and the difference between I_{BROJA} and I_{pool} are obvious. The definitions have exactly the same form as Eqs. (10)–(12), but the pooling distribution is replaced by the "worst" [highest-entropy] possible joint distribution in all these definitions. One consequence is that the redundant information here too is defined in terms of four different types of probability distributions, not just two.

We may also note that various inequalities derived in [3] (see Lemma 3, in particular) do not apply to our measures, because the assumption underlying that Lemma is that such measures are derived from joint distributions consistent with the marginals. The pooling distribution is not necessarily consistent with the marginals, as it forms a compromise instead. For example, for the simple "geometric average" pooling rule mentioned below, in Eq. (21), averaging over variable B would give us the square root of the probability distribution over A, renormalized.

3. Nonnegativity of information

The original idea of the PID was to decompose mutual information into nonnegative quantities, all interpretable as information. Taking information as a reduction in uncertainty (entropy) that accompanies going from initial to final probability distributions for the target variable $\mathcal{T}$, we saw that quite generically one of the four quantities introduced in (9) is not information. We also saw that unique information generically is indeed a difference between two entropies. As such, a requirement on unique information is that it be nonnegative. (In our case, synergistic information is automatically non-negative).

For our definition of unique information (11) the question of nonnegativity boils down to the question whether Alice and Bob can pool in such a way as to ensure that their fine is not more than either Alice's or Bob's individual fine. And indeed, they can, rather trivially: they could choose to report either always Alice's distribution or always Bob's, whichever one has the lowest uncertainty. Thus, one (easily met) requirement on the pooled distribution $\mathrm{P}_{\{A\},\{B\}}$ is that its expected uncertainty always be less than or at worst equal to the minimum of $H_{\{A\}}$ and $H_{\{B\}}$. That is, we require

$$H_{\{A\},\{B\}} \leqslant \min(H_{\{A\}}, H_{\{B\}}). \tag{18}$$

Referring back to the very simple way of pooling by taking the average distribution, that simple method does not fulfill this criterion, as is easily checked.

In this context, given that our definition of redundant information is not, strictly speaking, information, we do not require it to be nonnegative. In fact, as we will see below in the examples section, it can be negative for certain choices of pooling, and then it could be made positive *only* by increasing the uncertainty in the pooled distribution, since the other three entropies and probability distributions featuring in (12) are fixed. That ad-hoc fix, though, is counter to the meaning of synergy: the whole would be more than the sum of its parts, but only because we do not do our best to get as much information as we can from the parts.

For two source variables A and B it may well be that the intuitions behind unique, redundant, and synergistic information (as defined within PID) are incompatible. That is, whenever one defines one of the three quantities, the remaining two are fixed, but that particular way of fixing their magnitudes may not be in agreement with what the remaining two terms are supposed to mean. The idea that not all quantities are expressible as a difference between two entropies may be taken as an indication in that direction as well.

D. Pooling-based lattice

The original work on PID [1] argued for a particular lattice underlying the PID structure. That is, the concept of redundant information naturally leads to a partial order, as well as that of a unique least upper bound and a unique greatest lower bound that can be assigned to every pair of elements. One idea behind the construction of a lattice, formulated in terms of our experts from Sec. II, is as follows: if we have a collection of sets of experts, and within that collection, one set is a subset of another, the redundant information is already present in the subset, and so we can delete the superset. The lattice of collections that thus remains is the redundancy-based lattice.

Here we use pooling instead of redundancy as our fundamental notion, and this leads to the opposite approach: if within a collection of sets of experts, one set is a subset of another, then we should use the superset when we pool, not the subset. So we delete the subset from our collection. The collections remaining form a pooling-based lattice.

For example, for three variables (or experts) neither lattice contains the collection $\{A, B\}$, $\{A, C\}$, $\{B\}$. In the redundancy-based lattice, the set $\{A, B\}$ is removed; in the pooling-based lattice, the set $\{B\}$ is removed.

For two elements in the pooling-based lattice, for example, $\{A\}$ and $\{B, C\}$, one defines the least upper bound and the greatest lower bound as follows. The least upper bound is the collection $\{A\}$, $\{B, C\}$, which corresponds to pooling the probability distributions from Alice with that of Bob and Charlie jointly. The information thus obtained is larger or equal to that of the maximum obtainable from either Alice, or by Bob and Charlie jointly. The greatest lower bound is the empty collection $\varnothing$. The information obtainable from Alice, and from Bob and Charlie jointly, are both equal to or more than that obtainable from not consulting any expert. There is no larger collection of experts with that property.

With every node of the lattice we associate probability distributions (over $\mathcal{T}$, one distribution for each of the values the other variables in the collection may take) and a number, namely their average entropy. The empty set is the least element in the pooling-based lattice, and we associate the distribution $\mathrm{P}_{\varnothing}(\mathcal{T})$ and its entropy $H_{\varnothing}$ with it. References [10,36] derive the same lattice from different considerations.

III. POOLING PROBABILITY DISTRIBUTIONS

For two variables the PID depends on the probability distributions $P_{\{a,b\}}(\mathcal{T})$ two experts, Alice and Bob, agree to use if all they know are the marginal distributions and the corresponding conditional distributions $P_A(\mathcal{T}) = P(\mathcal{T}|A)$ and $P_B(\mathcal{T}) = P(\mathcal{T}|B)$.

There are several options, some involving optimizations, others simpler. In most examples a simple geometric averaging procedure is used to define the pooled distribution. When that procedure fails a more general one-parameter family of pooled distributions is used and optimized over that one parameter.

A. Minimum goal

We consider here a simple example where some previous measures of synergy disagree. But let us first see what the essence is of an even simpler example where there is consensus on synergistic information. If A and B are random bits and $\mathcal{T} = A \oplus B$, then everyone agrees there is only one bit of synergistic information present. The reason is simple, neither $\mathrm{P}_A(\mathcal{T})$ nor $\mathrm{P}_B(\mathcal{T})$, which are identical, contains any information, and only the joint distribution yields the full one bit of information through $\mathrm{P}(\mathcal{T}|A, B)$.

In contrast, consider the example from Table I. The two variables A, B are binary, the target variable is ternary, and there are only three possible combinations of values with

TABLE I. Simple example (5A from [5]) in which the marginal distributions $\mathrm{P}(\mathcal{T}, A)$ and $\mathrm{P}(\mathcal{T}, B)$ together contain as much information about the target variable as does the joint distribution $\mathrm{P}(\mathcal{T}, A, B)$. In this case, the BROJA measure agrees exactly (but via a different mechanism, see Sec. II C 2) with the pooling measure. In particular, there is no synergistic information.

$A, B, \mathcal{T}$	Probability
0,0,0	1/3
0,1,1	1/3
1,0,2	1/3

I_∂	I_{BROJA}	I_{ccs}	$I_{\min}$	I_{pool}
syn $A\&B$	0	0.138	0.333	0
unq $A\backslash B$	0.667	0.528	0.333	0.667
unq $B\backslash A$	0.667	0.528	0.333	0.667
red $A\&B$	0.252	0.390	0.585	0.252

nonzero probability. The interesting (debatable) situation occurs when both A and B have the value zero. We are certain in this case of the value of $\mathcal{T}$, even without knowing the joint distribution. This is because each variable eliminates one possibility, leaving just one value of $\mathcal{T}$ to occur with 100% probability. There is no synergistic information here since the joint distribution $P(A, B, \mathcal{T})$ cannot give us more information than we can obtain from the marginals $P(A, \mathcal{T})$ and $P(B, \mathcal{T})$.

Some measures (in particular, from those measures that we compare to later on in Sec. IV, we find that $I_{\min}$ from [1] and I_{ccs} from [5]) ascribe a nonzero amount of synergistic information to this case. On the other hand, the BROJA measure (mentioned above) does yield zero synergistic information.

What the simple example means for our pooling rules is twofold:

(1) If for some combination of the values of the variables all marginal distributions are identical, then the pooled distribution for that combination will have to equal that distribution.

(2) If for a particular combination of variables one of the marginal distributions indicates a certain value of $\mathcal{T}$ appears with 0% probability, then this should be true for the pooled distribution as well.

These are taken as two necessary (although by no means sufficient) requirements on pooling, in addition to the requirement (18) we found above.

B. Simple pooling rules

There are a few different rules that have been proposed for pooling expert opinions, with different circumstances leading to different conclusions about which way is appropriate. (See Refs. [27–30] for background information about this topic).

Here are two ways of pooling that seem inapplicable to our specific case, where each expert opinion $\mathrm{P}(\mathcal{T}|X)$ for $X = A, B, C, \ldots$ is derivable from a joint distribution $\mathrm{P}(\mathcal{T}, A, B, C, \ldots)$. First, there is the (weighted) average

$$P(\mathcal{T}|A, B, C, \ldots) = \sum_X w_X \mathrm{P}(\mathcal{T}|X), \qquad (19)$$

with $w_X > 0$ and $\sum_X w_X = 1$. This may be correct when experts may contradict each other and a compromise is needed. But in our case it fails the second rule.

Another rule would be to take the product,

$$P(\mathcal{T}|A, B, C, \ldots) = \Pi_X \mathrm{P}(\mathcal{T}|X). \qquad (20)$$

This would apply to a case where we try to estimate the average value of $\mathcal{T}$ and we have independent data that, nonetheless, do estimate the same average. Then it is true that the more data we have the better our average should be determined. The variance in the distribution of possible values for the average should decrease with more data. But this rule violates our first requirement on pooling.

1. *Geometric average*

The third rule does fit both our pooling requirements and, moreover, has other advantages, which, however, do not concern us here [27–30]. Although the following definition used here can easily be extended to more than two probability distributions (or agents) we focus here on the case of two agents, Alice and Bob, first. (See the Appendix about the extension to more than two experts). We define for each pair of values a, b for A, B the geometric average:

$$P_{\{A\}\{B\}}(\mathcal{T}|a, b) = \sqrt{\mathrm{P}(\mathcal{T}|a)\mathrm{P}(\mathcal{T}|b)}/Z_{ab}, \qquad (21)$$

where the normalization factor

$$Z_{ab} = \sum_t \sqrt{\mathrm{P}(t|a)\mathrm{P}(t|b)} \qquad (22)$$

is needed to define a proper probability distribution. Z_{ab} is an overlap between two distribution functions, and is in fact the Bhattacharyya measure [37]. It lies between zero (for orthogonal distributions, which have no common support) and one, for identical distributions.

The expected value of Alice's and Bob's fine is

$$H_{\{A\}\{B\}} = \tfrac{1}{2}H_{\{A\}} + \tfrac{1}{2}H_{\{B\}} - \mathcal{B}, \qquad (23)$$

with the nonnegative quantity $\mathcal{B}$ given by

$$\mathcal{B} = \sum_a \sum_b \mathrm{P}(a, b)\log(1/Z_{ab}), \qquad (24)$$

where

$$\mathrm{P}(a, b) = \sum_t \mathrm{P}(t, a, b) \qquad (25)$$

is the joint distribution for a, b. Note that an average fine equal to the sum of the first two terms can be obtained by the agents simply by reporting either P_A or P_B with 50% probability. That is why we may consider $\mathcal{B}$ as a "bonus," extra money saved by Alice and Bob when they pool their probability distributions instead of randomly choosing P_A or P_B.

Note that Alice and Bob cannot determine the value of the bonus $\mathcal{B}$ as long as they do not know the joint distribution $\mathrm{P}(a, b)$. They could still determine the *minimum* bonus by minimizing over all possible joint distributions $\mathrm{P}_{\mathrm{poss}}(a, b)$ consistent with $\mathrm{P}(a)$ and $\mathrm{P}(b)$. That is one way for them to see if the bonus is large enough to satisfy the inequality (18), i.e., whether

$$\mathcal{B} \geqslant \tfrac{1}{2}|H_{\{A\}} - H_{\{B\}}|. \qquad (26)$$

If one needs to define a PID for a system for which one actually knows all probability distributions, one could be satisfied with requiring (26) to hold for the actual distribution $\mathrm{P}(a, b)$.

Now that we have an expression for $H_{\{A\}\{B\}}$ we can write the following relation between unique information and the bonus,

$$I_{\mathrm{unq}\,A\backslash B} + I_{\mathrm{unq}\,B\backslash A} = 2\mathcal{B}, \tag{27}$$

which neatly quantifies the intuition that unique information is determined by the extent to which the probability distributions $\mathrm{P}_{\{A\}}$ and $\mathrm{P}_{\{B\}}$ differ.

The individual unique information quantities are given by

$$\begin{aligned} I_{\mathrm{unq}\,A\backslash B} &= \mathcal{B} + \tfrac{1}{2}(H_{\{B\}} - H_{\{A\}}), \\ I_{\mathrm{unq}\,B\backslash A} &= \mathcal{B} + \tfrac{1}{2}(H_{\{A\}} - H_{\{B\}}), \end{aligned} \tag{28}$$

which are nonnegative thanks to the requirement (26).

2. *One-parameter family*

This method of pooling is called "logarithmic" and generalizes easily to using pooling distributions of the form

$$P^{(w)}_{\{A\}\{B\}}(\mathcal{T}|a, b) = \mathrm{P}(\mathcal{T}|a)^{w}\mathrm{P}(\mathcal{T}|b)^{1-w}/Z_{w,ab}, \tag{29}$$

for $0 \leqslant w \leqslant 1$, with normalization

$$Z_{w,ab} = \sum_t \mathrm{P}(t|a)^{w}\mathrm{P}(t|b)^{1-w}. \tag{30}$$

The expected fine can be written as

$$H_{\{A\}\{B\}} = wH_{\{A\}} + (1-w)H_{\{B\}} - \mathcal{B}_w \tag{31}$$

with

$$\mathcal{B}_w = \sum_a\sum_b \mathrm{P}(a, b)\log(1/Z_{w,ab}), \tag{32}$$

which is nonnegative, as can be shown using the Rogers-Hölder's inequality [38]. (For the simpler case of $w = 1/2$ adopted above, this inequality reduces to the more straightforward Cauchy-Schwarz inequality),

In general, we recommend using the one-parameter family (29) of pooling distributions if the simple geometric average ($w = 1/2$) violates condition (18). One way to choose the weight w would be to make it dependent on the entropies $H_{\{A\}}$ and $H_{\{B\}}$. For example, one might choose $w = 2^{-H_{\{A\}}}/(2^{-H_{\{A\}}} + 2^{-H_{\{B\}}})$, thus giving more weight to the lower-entropy distribution. Another way is discussed next. It is important to note that there is always a value of w such that all our requirements on the pooling distribution are met, namely those listed at the end of Sec. III A and (18).

C. Optimized pooling

Denote the set of possible joint distributions $P_{\mathrm{poss}}(\mathcal{T}, A, B)$ consistent with P_A and P_B as Δ_P. Denote the set of all joint probability distributions by Ω. That is, $\Delta_P \subset \Omega$. If Alice and Bob choose a distribution $\omega(\mathcal{T}, A, B)$ from Ω, then, relative to a possible distribution $P_{\mathrm{poss}}(\mathcal{T}, A, B)$ their expected fine is given by

$$F(\omega, \mathrm{P}_{\mathrm{poss}}) := -\sum_a\sum_b\sum_t P_{\mathrm{poss}}(t, a, b)\log\frac{\omega(t, a, b)}{\omega(a, b)}$$

with $\omega(a, b) = \sum_t \omega(t, a, b)$. They would like to minimize their fine, although they do not know which possible distribution $\mathrm{P}_{\mathrm{poss}} \in \Delta_P$ they have. They could find the worst-case distribution $\mathrm{P}_{\mathrm{poss}}$ for each choice of ω, and then minimize over all possible choices of ω. Their "best" fine would then be

$$F_{\mathrm{best}} = \min_{\omega\in\Omega}\max_{\mathrm{P}_{\mathrm{poss}}\in\Delta_P} F(\omega, \mathrm{P}_{\mathrm{poss}}). \tag{33}$$

TABLE II. Comparison of a few PIDs for the AND example. I_{pool} is close to I_{dep}, and to a lesser degree to I_{ccs}, but clearly disagrees with the BROJA measure. See Table 5 from Ref. [9].

$A, B, \mathcal{T}$	Probability
0,0,0	1/4
0,1,0	1/4
1,0,0	1/4
1,1,1	1/4

I_∂	I_{BROJA}	I_{ccs}	I_{dep}	I_{pool}
syn $A\&B$	0.500	0.292	0.270	0.250
unq $A\backslash B$	0	0.208	0.230	0.250
unq $B\backslash A$	0	0.208	0.230	0.250
red $A\&B$	0.311	0.104	0.082	0.061

A simpler version of this idea would make use of a smaller subset of all distributions ω. For example, they may restrict to the type of logarithmic pooling distributions discussed above, but with arbitrary weights:

$$\frac{\omega(t, a, b)}{\omega(a, b)} = \mathrm{P}(t|a)^{w}\mathrm{P}(t|b)^{1-w} \tag{34}$$

for all $0 \leqslant w \leqslant 1$, and then perform the minimization just over the parameter w. In the next ection with examples, we will display one case where $w = 1/2$ violates the requirement (26). That is, the pooling distribution $\sqrt{\mathrm{P}(\mathcal{T}|a)\mathrm{P}(\mathcal{T}|b)}/Z_{ab}$ is inferior to the better one of $\mathrm{P}_{\{A\}}$ and $\mathrm{P}_{\{B\}}$. But minimization over w leads to a pooling distribution of the form (29) superior to both $\mathrm{P}_{\{A\}}$ and $\mathrm{P}_{\{B\}}$, which does satisfy all our requirements.

IV. EXAMPLES

In the following, I_{pool} is based on logarithmic pooling with weight $w = 1/2$, unless stated otherwise.

A. Standard examples

Many examples of distributions have been tested on many different PIDs. Here we just use a small selection of three examples, on a small selection of PIDs to compare the PID that results from the simplest pooling strategy. The results for other measures used here are all conveniently found in Table 5 from Ref. [9]. One point of that table was to show how various measures clearly disagree with the original measure I_{min} from [1], and it turns out the same is true for I_{pool}. For the TWO BIT example, I_{pool} agrees with all three measures, and so is not displayed here. The comparisons for the NOT TWO example are similar to those for AND, and are for that reason not displayed here either.

TABLE III. Comparison of a few PIDs for the DIFF example. I_{pool} agrees exactly with I_{dep} and I_{BROJA}, but disagrees with I_{ccs}. See Table 5 from Ref. [9].

$A, B, \mathcal{T}$	Probability
0,0,0	1/4
0,0,1	1/4
0,1,0	1/4
1,0,1	1/4

I_∂	I_{BROJA}	I_{ccs}	I_{dep}	I_{pool}
syn $A\&B$	0	0.085	0	0
unq $A\backslash B$	0.189	0.104	0.189	0.189
unq $B\backslash A$	0.189	0.104	0.189	0.189
red $A\&B$	0.123	0.208	0.123	0.123

The three examples we do consider are in Tables II, III, and IV. One point here is to show that there is always an example where I_{pool} disagrees with a given measure. On the other hand, for those same examples there also are other measures I_{pool} agrees with, either exactly or approximately.

B. When the geometric average fails

For the example of Table V (found numerically, then simplified to make all percentages integers) the pooling distribution obtained by taking the geometric average of Alice's and Bob's individual distributions fails to meet requirement (26) or, equivalently, (18).

That is, Alice and Bob would be better off (i.e., incur a smaller fine on average) simply always reporting the more informative individual distribution, in this case Bob's [with an average fine (entropy) of 0.321 bits vs 0.472 bits for Alice's]. They would be aware of the flaw in the simple pooling method, and so could indeed report Bob's distribution $P_{\{B\}}(\mathcal{T})$. In that case, the unique information from Alice would be identically zero (as she does not contribute anything to the pooling distribution).

However, by considering the logarithmic pooling rule with arbitrary unequal weights, they would find that by using a weight $w \approx 0.12$ for Alice and hence a weight $1 - w \approx 0.88$ for Bob, they would do even better, as displayed in the Table V. Alice would thus contribute a very small but nonzero amount of unique information.

TABLE IV. Comparison of a few PIDs for the PNT. UNQ. example. I_{pool} agrees with I_{ccs}. See Table 5 from Ref. [9].

$A, B, \mathcal{T}$	Probability
0,1,1	1/4
1,0,1	1/4
0,2,2	1/4
2,0,2	1/4

I_∂	I_{BROJA}	I_{ccs}	I_{dep}	I_{pool}
syn $A\&B$	0.500	0	0.250	0
unq $A\backslash B$	0	0.500	0.250	0.500
unq $B\backslash A$	0	0.500	0.250	0.500
red $A\&B$	0.500	0	0.250	0

TABLE V. Comparison of two different logarithmic pooling distributions.

$A, B, \mathcal{T}$	Probability
0,0,0	0.50
1,0,0	0.39
1,1,0	0.01
0,0,1	0.04
0,1,1	0.04
1,0,1	0.01
1,1,1	0.01

I_∂	I_{pool}, $w = 1/2$	I_{pool}, $w = 0.12$
syn $A\&B$	0.051	0.026
unq $A\backslash B$	−0.023	0.002
unq $B\backslash A$	0.108	0.133
red $A\&B$	0.040	0.015

C. Negative $\Delta I_{\mathrm{red}\,A\&B}$

Table VI presents an example where the simple pooling method leads to negative redundant information, while still meeting requirement (26) or, equivalently, (18). As mentioned before, this can be seen as a consequence of that quantity depending on four different probability distributions, which means its interpretation as "information" is not straightforward. The example is called ReducedOr in [5] where it is attributed to Joseph Lizier. The point of that example was to locate a defect in the BROJA measure. That is, the example clearly contains unique information from both variables, even though the BROJA measure (as well as the original I_{min} measure from Ref. [1]) assigns zero unique information. In fact, when either variable takes on the value one, it provides the unique information that the target variable must have the value one.

Our pooling measure agrees with that verdict, but the main reason for displaying this example here, is that the shared (or redundant) information is negative. Recalling the expression for redundant information, we could make it equal to zero

TABLE VI. Comparison of a few PIDs for the ReducedOr example discussed in [5]. I_{pool} agrees with I_{ccs} against I_{BROJA} that there is nonzero unique information in this case.

$A, B, \mathcal{T}$	Probability
0,0,0	1/2
1,0,1	1/4
0,1,1	1/4

I_∂	I_{BROJA}	I_{ccs}	I_{pool}
syn $A\&B$	0.69	0.38	0.29
unq $A\backslash B$	0	0.31	0.40
unq $B\backslash A$	0	0.31	0.40
red $A\&B$	0.31	0	−0.09

only in an ad-hoc manner, by increasing the entropy of the pooled distribution by 0.09 units (and then all four quantities would numerically agree with $I_{\rm ccs}$, as is easily checked). Since increasing that uncertainty would correspond to Alice and Bob increasing their fine, this would go against the spirit of pooling. In any case, as noted several times, the fact that $\Delta I_{{\rm red}\,A\&B}$ is not simply a difference between two entropies, and is thus not information, also (arguably) eliminates the requirement for it to be nonnegative. We thus accept here the result displayed in Table VI.

V. CONCLUSIONS

Within the context of the partial information decomposition (PID, [1]), synergistic information is meant to quantify how much "the whole" is more than "the sum of its parts." In the case of two source variables A and B that provide information about a target variable $\mathcal{T}$, we identified here "the parts" with marginal probability distributions $\mathrm{P}(A,\mathcal{T})$ and $\mathrm{P}(B,\mathcal{T})$ and "the whole" with the joint probability distribution $\mathrm{P}(A,B,\mathcal{T})$. Synergistic information is then the average reduction in uncertainty (as measured by entropy) upon using the joint distribution rather than a specific distribution aggregated (or pooled) from the two marginal distributions for making predictions about $\mathcal{T}$.

This idea of pooling [27] was shown to lead to a lattice underlying the PID, which differs from the original redundancy-based lattice used by Williams and Beer in their original work proposing the PID [1]. Each element of the lattice is a collection of sets of the variables, with no set in the collection being a subset of another. With each element of the lattice we can associate pooled probability distributions, their average entropy, and synergistic information. For example, given the collection consisting of $\{A,B\}$ and $\{C\}$, we combine $\mathrm{P}(\mathcal{T}|A,B)$ and $\mathrm{P}(\mathcal{T}|C)$ into pooled distributions, and calculate their average entropy (averaged over all variables $A,B,C,\mathcal{T}$). The difference between the average entropy of the pooled distributions and the entropy of the full distribution $\mathrm{P}(\mathcal{T}|A,B,C)$ is defined to equal the synergistic information associated with that element of the lattice.

We considered logarithmic pooling as a simple and convenient pooling method which can be optimized to provide the "best" logarithmic way to pool information. It provides us with sensible definitions of synergistic information as well as of unique information, both satisfying the basic definition of information as a reduction in uncertainty when switching from one type of probability distribution to a better one. Inevitably, "redundant information" then involves four different types of probability distributions and thus is not of that basic form. The possible pooling distributions (29) are parametrized by just a single parameter w, and the only requirement on w is that the condition (18) be fulfilled. That condition is equivalent to requiring unique information to be nonnegative. There is always such a value for w.

Other (more complicated) ways of pooling distributions are possible, and depending on one's definition of the "best" way to do this, one finds different measures of synergistic information and hence, of unique information. This freedom of choice illustrates the ambiguity in the definition of a PID.

APPENDIX: MORE THAN TWO SOURCES

The idea of logarithmic pooling is easily extended to more than two experts, as follows. For experts $A,B,C,\dots$ we may define

$$\mathrm{P}_{\{A\},\{B\},\{C\},\dots}(\mathcal{T})=\mathrm{P}_A(\mathcal{T})^{w_A}\mathrm{P}_B(\mathcal{T})^{w_B}\mathrm{P}_C(\mathcal{T})^{w_C}\dots, \tag{A1}$$

where the positive weights $w_A,w_B,w_C,\dots$ add up to unity. For three experts one may wonder whether to choose the weights equal, as we did for two experts by default. There is a good reason not to always do this: once we have calculated the redundant information shared amongst pairs of experts, we may then penalize such redundancy by lowering the sum of the weights for such pairs. Thus, if Alice and Bob's information is purely shared, and Charlie's information is unique relative to Alice and Bob, then it makes sense to set $w_C=1/2$ and $w_A=w_B=1/4$, for example. That is, Alice and Bob together are assigned the same weight as Charlie. Since this case provides one extreme, we may set the general rule for three experts:

$$\tfrac{1}{4}\leqslant w_k\leqslant\tfrac{1}{2}, \tag{A2}$$

where within this range there is freedom depending on how one takes into account the pairwise redundant and unique information.

For more than three experts (say, N_e of them) it becomes more complicated to design general rules for assigning weights. The upper bound of $1/2$ still holds (and would be correct only if all experts but one provide only shared information: that one expert gets the weight $1/2$; the other N_e-1 experts share equally the remaining weight) and we may thus generalize (A2) as

$$\frac{1}{2(N_e-1)}\leqslant w_k\leqslant\frac{1}{2}. \tag{A3}$$

The idea of pooling for more than two experts still leads to a lattice, as explained before in Sec. II D.

[1] P. L. Williams and R. D. Beer, Nonnegative decomposition of multivariate information, arXiv:1004.2515.

[2] M. Harder, C. Salge, and D. Polani, Bivariate measure of redundant information, Phys. Rev. E **87**, 012130 (2013).

[3] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, Quantifying unique information, Entropy **16**, 2161 (2014).

[4] V. Griffith and C. Koch, Quantifying synergistic mutual information, in *Guided Self-Organization: Inception* (Springer, 2014), pp. 159–190.

[5] R. A. A. Ince, Measuring multivariate redundant information with pointwise common change in surprisal, Entropy **19**, 318 (2017).

[6] J. Rauh, Secret sharing and shared information, Entropy **19**, 601 (2017).

[7] C. Finn and J. T. Lizier, Pointwise partial information decompositionusing the specificity and ambiguity lattices, Entropy **20**, 297 (2018).

[8] R. G. James, J. Emenheiser, and J. P. Crutchfield, Unique information via dependency constraints, J. Phys. A: Math. Theor. **52**, 014002 (2018).

[9] R. G. James, J. Emenheiser, and J. P. Crutchfield, Unique information and secret key agreement, Entropy **21**, 12 (2018).

[10] N. Ay, D. Polani, and N. Virgo, Information decomposition based on cooperative game theory, arXiv:1910.05979.

[11] R. G. James and J. P. Crutchfield, Multivariate dependence beyond shannon information, Entropy **19**, 531 (2017).

[12] F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. F. de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno *et al.*, The physics of higher-order interactions in complex systems, Nat. Phys. **17**, 1093 (2021).

[13] F. E. Rosas, P. A. M. Mediano, M. Gastpar, and H. J. Jensen, Quantifying high-order interdependencies via multivariate extensions of the mutual information, Phys. Rev. E **100**, 032305 (2019).

[14] T. M. S. Tax, P. A. M. Mediano, and M. Shanahan, The partial information decomposition of generative neural network models, Entropy **19**, 474 (2017).

[15] M. Wibral, C. Finn, P. Wollstadt, J. T. Lizier, and V. Priesemann, Quantifying information modification in developing neural networks via partial information decomposition, Entropy **19**, 494 (2017).

[16] M. Wibral, V. Priesemann, J. W. Kay, J. T. Lizier, and W. A. Phillips, Partial information decomposition as a unified approach to the specification of neural goal functions, Brain Cog. **112**, 25 (2017).

[17] T. F. Varley and E. Hoel, Emergence as the conversion of information: a unifying theory, Phil. Trans. R. Soc. A **380**, 20210150 (2022).

[18] P. A. M. Mediano, F. E. Rosas, A. I. Luppi, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor, Greater than the parts: a review of the information decomposition approach to causal emergence, Phil. Trans. R. Soc. A **380**, 20210246 (2022).

[19] A. I. Luppi, P. A. M. Mediano, F. E. Rosas, J. Allanson, J. D. Pickard, R. L. Carhart-Harris, G. B. Williams, M. M. Craig, P. Finoia, A. M. Owen *et al.*, A synergistic workspace for human consciousness revealed by integrated information decomposition, BioRxiv (2020), doi: 10.1101/2020.11.25.398081.

[20] J. T. Lizier, N. Bertschinger, J. Jost, and M. Wibral, Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work, Entropy **20**, 307 (2018).

[21] A. Kolchinsky, A novel approach to the partial information decomposition, Entropy **24**, 403 (2022).

[22] E. Schneidman, W. Bialek, and M. J. Berry, Synergy, redundancy, and independence in population codes, J. Neurosci. **23**, 11539 (2003).

[23] P. E. Latham and S. Nirenberg, Synergy, redundancy, and independence in population codes, revisited, J. Neurosci. **25**, 5195 (2005).

[24] A. J. Gutknecht, M. Wibral, and A. Makkeh, Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic, Proc. R. Soc. A **477**, 20210110 (2021).

[25] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).

[26] J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 1999).

[27] N. C. Lind and A. S. Nowak, Pooling expert opinions on probability distributions, J. Eng. Mech. **114**, 328 (1988).

[28] D. Poole and A. E. Raftery, Inference for deterministic simulation models: the bayesian melding approach, J. Am. Stat. Assoc. **95**, 1244 (2000).

[29] L. M. Carvalho, D. A. M. Villela, F. C. Coelho, and L. S. Bastos, Bayesian inference for the weights in logarithmic pooling, Bayesian Analysis **18**, 223 (2023).

[30] E. Neyman and T. Roughgarden, No-regret learning with unbounded losses: The case of logarithmic pooling, arXiv:2202.11219.

[31] G. A. Gratzer, *Lattice Theory: Foundation* (Springer, Basel: Birkhäuser, 2011), Vol. 2.

[32] For the goal of defining the PID it is assumed that these experimentally estimated conditional probability distributions are accurate to an arbitrary degree. Effects of statistical fluctuations in estimates of probability distributions for which the PID is to be determined seem not to have been considered.

[33] T. Gneiting and A. E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc. **102**, 359 (2007).

[34] This is a strong assumption, which would not apply to opinions held by meteorologists about the weather or by economists about inflation, unless they all used the same weather or inflation model, respectively. It does hold for the typical models for which the PID is meant to be used.

[35] A. J. Bell, The co-information lattice, in *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA* (2003), Vol. 2003, pp. 921–926.

[36] F. E. Rosas, P. A. M. Mediano, B. Rassouli, and A. B. Barrett, An operational information decomposition via synergistic disclosure, J. Phys. A: Math. Theor. **53**, 485001 (2020).

[37] A. Bhattacharyya, On a measure of divergence between two multinomial populations, in *Sankhyā: The Indian Journal of Statistics* (1946), pp. 401–406.

[38] L. Maligranda, Why hölder's inequality should be called rogers' inequality, Math. Inequalities Appl. **1**, 69 (1998).