entropy

MDPI

Article

Understanding Interdependency Through Complex Information Sharing

Fernando Rosas [1,*], Vasilis Ntranos [2], Christopher J. Ellison [3], Sofie Pollin [1] and Marian Verhelst [1]

1 Departement Elektrotechniek, KU Leuven, Leuven 3001, Belgium; sofie.pollin@esat.kuleuven.be (S.P.); marian.verhelst@esat.kuleuven.be (M.V.)
2 Department of Electrical Engineering & Computer Sciences, UC Berkeley, Berkeley, CA 94720, USA; ntranos@usc.edu
3 Center for Complexity and Collective Computation, University of Wisconsin-Madison, Madison, WI 53706, USA; cellison@wisc.edu
* Correspondence: fernando.rosas@esat.kuleuven.be; Tel.: +32-16-32-89-81

Academic Editor: Raúl Alcaraz Martínez
Received: 14 September 2015; Accepted: 22 December 2015; Published: 26 January 2016

Abstract: The interactions between three or more random variables are often nontrivial, poorly understood and, yet, are paramount for future advances in fields such as network information theory, neuroscience and genetics. In this work, we analyze these interactions as different modes of information sharing. Towards this end, and in contrast to most of the literature that focuses on analyzing the mutual information, we introduce an axiomatic framework for decomposing the joint entropy that characterizes the various ways in which random variables can share information. Our framework distinguishes between interdependencies where the information is shared redundantly and synergistic interdependencies where the sharing structure exists in the whole, but not between the parts. The key contribution of our approach is to focus on symmetric properties of this sharing, which do not depend on a specific point of view for differentiating roles between its components. We show that our axioms determine unique formulas for all of the terms of the proposed decomposition for systems of three variables in several cases of interest. Moreover, we show how these results can be applied to several network information theory problems, providing a more intuitive understanding of their fundamental limits.

Keywords: Shannon information; multivariate dependencies; mutual information; synergy; information decomposition; shared information

1. Introduction

Interdependence is a key concept for understanding the rich structures that can be exhibited by biological, economic and social systems [1,2]. Although this phenomenon lies at the heart of our modern interconnected world, there is still no solid quantitative framework for analyzing complex interdependencies, this being crucial for future advances in a number of disciplines. In neuroscience, researchers desire to identify how various neurons affect an organism's overall behavior, asking to what extent the different neurons are providing redundant or synergistic signals [3]. In genetics, the interactions and roles of multiple genes with respect to phenotypic phenomena are studied, e.g., by comparing results from single and double knockout experiments [4]. In graph and network theory, researchers are looking for measures of the information encoded in node interactions in order to quantify the complexity of the network [5]. In communication theory, sensor networks usually generate strongly-correlated data [6]; a haphazard design might not account for these interdependencies and, undesirably, will process and transmit redundant information across the network, degrading the efficiency of the system.

The dependencies that can exist between two variables have been extensively studied, generating a variety of techniques that range from statistical inference [7] to information theory [8]. Most of these approaches require to differentiate the role of the variables, e.g., between a *target* and a *predictor*. However, the extension of these approaches to three or more variables is not straightforward, as a binary splitting is, in general, not enough to characterize the rich interplay that can exist between variables. Moreover, the development of more adequate frameworks has been difficult, as most of our theoretical tools are rooted in sequential reasoning, which is adept at representing linear flows of influences, but is not as well-suited for describing distributed systems or complex interdependencies [9].

In this work, our approach is to understand interdependencies between variables as *information sharing*. In the case of two variables, the portion of the variability that can be predicted corresponds to information that a target and a predictor have in common. Following this intuition, we present a framework that decomposes the total information of a distribution according to how it is shared among the variables. Our framework is novel in combining the hierarchical decomposition of higher-order interactions, as developed in [10], with the notion of synergistic information, as proposed in [11]. In contrast to [10], we study the information that exists in the system itself without comparing it to other related distributions. In contrast to [11], we analyze the joint entropy instead of the mutual information, looking for symmetric properties of the system. Note that a different approach for relating the tools presented in [10] and the idea of synergistic information has been presented independently in [12].

One important contribution of this paper is to distinguish *shared information* from *predictability*. Predictability is a concept that requires a bipartite system divided into predictors and targets. As different splittings of the same system often yield different conclusions, we see predictability as a directed notion that strongly depends on one's "point of view". In contrast, we see shared information as a property of the system itself, which does not require differentiated roles between its components. Although it is not possible in general to find a unique measure of predictability, we show that the shared information can be uniquely defined for systems of three variables in a number of interesting scenarios.

Additionally, our framework provides new insight into various problems of network information theory. Interestingly, many of the problems of network information theory that have been solved are related to systems that present a simple structure in terms of shared information and synergies, while most of the open problems possess a more complex mixture of them.

The rest of this article is structured as follows. First, Section 2 introduces the notions of the hierarchical decomposition of dependencies and synergistic information, reviewing the state of the art and providing the necessary background for the unfamiliar reader. Section 3 presents our axiomatic decomposition for the joint entropy, focusing on the fundamental case of three random variables. Then, we illustrate the application of our framework for various cases of interest: pairwise independent variables in Section 4, pairwise maximum entropy distributions and Markov chains in Section 5 and multivariate Gaussians in 6. After that, Section 7 presents the first application of this framework in settings of fundamental importance for network information theory. Finally, Section 8 summarizes our conclusions.

2. Preliminaries and the State of the Art

One way of analyzing the interactions between the random variables $\mathbf{X} = (X_1, \ldots, X_N)$ is to study the properties of the correlation matrix $\mathcal{R}_{\mathbf{X}} = \mathbb{E}\{\mathbf{X}\mathbf{X}^t\}$. However, this approach only captures linear relationships, and hence, the picture provided by $\mathcal{R}_{\mathbf{X}}$ is incomplete. Another possibility is to study the matrix $\mathcal{I}_{\mathbf{X}} = [I(X_i; X_j)]_{i,j}$ of mutual information terms. This matrix captures the existence of both linear and nonlinear dependencies [13], but its scope is restricted to pairwise relationships and, thus, misses all higher-order structure. To see an example of how this can happen, consider two independent fair coins X_1 and X_2, and let $X_3 := X_1 \,\texttt{XOR}\, X_2$ be the output of an *exclusive-or* logic

gate. The mutual information matrix $\mathcal{I}_{\mathbf{X}}$ has all its off-diagonal elements equal to zero, making it indistinguishable from an alternative situation where X_3 is just another independent fair coin.

For the case of $\mathcal{R}_{\mathbf{X}}$, a possible next step would be to consider higher-order moment matrices, such as co-skewness and co-kurtosis. We seek their information-theoretic analogs, which complement the description provided by $\mathcal{I}_{\mathbf{X}}$. One method of doing this is by studying the information contained in marginal distributions of increasingly larger sizes; this approach is presented in Section 2.1. Other methods try to provide a direct representation of the information that is shared between the random variables; they are discussed in Sections 2.2–2.4.

2.1. Negentropy and Total Correlation

When the random variables that compose a system are independent, their joint distribution is given by the product of their marginal distributions. In this case, the marginals contain all that is to be learned about the statistics of the entire system. For an arbitrary joint probability density function (pdf), knowing the single variable marginal distributions is not enough to capture all there is to know about the statistics of the system.

To quantify this idea, let us consider the information stored in N discrete random variables $\mathbf{X} = (X_1, \ldots, X_N)$ with joint pdf $p_{\mathbf{X}}$, where each X_j takes values in a finite set with cardinality Ω_j. Here, we refer to a Bayesian interpretation of the Shannon information as described in [14], which corresponds to the state of knowledge that an observer has with respect to a system as described by its probability distribution; in this context, uncertainty in the system corresponds to information that can be extracted by performing measurements. The maximal amount of information that could be stored in $\mathbf{X}$ is $H^{(0)} = \sum_j \log \Omega_j$, which corresponds to the entropy of the pdf $p_{\mathbf{U}} := \prod_j \overline{p}_{X_j}$, where $\overline{p}_{X_j}(x) = 1/\Omega_j$ is the uniform distribution for each random variable X_j. On the other hand, the joint entropy $H(\mathbf{X})$ with respect to the true distribution $p_{\mathbf{X}}$ measures the actual uncertainty that the system possesses. Therefore, the difference

$$\mathcal{N}(\mathbf{X}) := H^{(0)} - H(\mathbf{X}) \tag{1}$$

corresponds to the decrease of the uncertainty about the system that occurs when one learns its pdf, *i.e.*, the information about the system that is contained in its statistics. This quantity is known as *negentropy* [15] and can also be computed as

$$\mathcal{N}(\mathbf{X}) = \sum_j [\log \Omega_j - H(X_j)] + \left(\sum_j H(X_j) - H(\mathbf{X}) \right) \tag{2}$$

$$= D\left(\prod_j p_{X_j} \,\middle|\middle|\, p_{\mathbf{U}} \right) + D\left(p_{\mathbf{X}} \,\middle|\middle|\, \prod_j p_{X_j} \right) \,, \tag{3}$$

where p_{X_j} is the marginal of the variable X_j and $D(\cdot||\cdot)$ is the Kullback–Leibler divergence. In this way, Equation (3) decomposes the negentropy into a term that corresponds to the information given by simple marginals and a term that involves higher-order marginals. The second term is known as the *total correlation* (TC) [16] (also known as *multi-information* [17]), which is equal to the mutual information for the case of $n = 2$. Because of this, the TC has been suggested as an extension of the notion of mutual information for multiple variables.

An elegant framework for decomposing the TC can be found in [10] (for an equivalent formulation that does not rely on information geometry, see [18]). Let us call k-marginals the distributions that are obtained by marginalizing the joint pdf over $N - k$ variables. Note that the k-marginals provide a more detailed description of the system than the $(k-1)$-marginals, as the latter can be directly computed from the former by marginalizing the corresponding variables. In the case where only the one-marginals are known, the simplest guess for the joint distribution is $\tilde{p}^{(1)}_{\mathbf{X}} = \prod_j p_{X_j}$. One way of generalizing this for the case where the k-marginals are known is by using the *maximum*

entropy principle [14], which suggests choosing the distribution that maximizes the joint entropy while satisfying the constrains given by the partial (k-marginal) knowledge. Let us denote by $\tilde{p}_{\mathbf{X}}^{(k)}$ the pdf that achieves the maximum entropy while being consistent with all of the k-marginals, and let $H^{(k)} = H(\{\tilde{p}_{\mathbf{X}}^{(k)}\})$ denote its entropy. Note that $H^{(k)} \geq H^{(k+1)}$, since the number of constrains that are involved in the maximization process that generates $H^{(k)}$ increases with k. It can therefore be shown that the following generalized Pythagorean relationship holds for the total correlation:

$$\mathrm{TC} = H^{(1)} - H(\mathbf{X}) = \sum_{k=2}^{N} \left[H^{(k-1)} - H^{(k)} \right] = \sum_{k=2}^{N} D\left(\tilde{p}_{\mathbf{X}}^{(k)} || \tilde{p}_{\mathbf{X}}^{(k-1)} \right) \coloneqq \sum_{k=2}^{N} \Delta H^{(k)} \ . \tag{4}$$

Above, $\Delta H^{(k)} \geq 0$ measures the additional information that is provided by the k-marginals that was not contained in the description of the system given by the $(k-1)$-marginals. In general, the information that is located in terms with higher values of k is due to dependencies between groups of variables that cannot be reduced to combinations of dependencies between smaller groups.

It has been observed that in many practical scenarios, most of the TC of the measured data is provided by the lower marginals. It is direct to see that

$$\mathrm{TC} - \sum_{k=2}^{k_0} \Delta H^{(k)} = \sum_{k=k_0+1}^{N} \Delta H^{(k)} = D\left(p_{\mathbf{X}} || \tilde{p}_{\mathbf{X}}^{(k_0)} \right) \ . \tag{5}$$

Therefore, if there exists a k_0, such that all $\Delta H^{(k)}$ are small for $k > k_0$, then $\tilde{p}_{\mathbf{X}}^{(k_0)}$ provides an accurate approximation for $p_{\mathbf{X}}$ from the point of view of the Kullback–Leibler divergence. In fact, it has been shown that pairwise maximum entropy models (*i.e.*, $k_0 = 2$) can provide an accurate description of the statistics of many biological systems [19–22] and also some social organizations [23,24].

2.2. Internal and External Decompositions

An alternative approach to study the interdependencies between many random variables is to analyze the ways in which they share information. This can be done by decomposing the joint entropy of the system. For the case of two variables, the joint entropy can be decomposed as

$$H(X_1, X_2) = I(X_1; X_2) + H(X_1|X_2) + H(X_2|X_1) \ , \tag{6}$$

suggesting that it can be divided into shared information, $I(X_1; X_2)$, and into terms that represent information that is exclusively located in a single variable, *i.e.*, $H(X_1|X_2)$ for X_1 and $H(X_2|X_1)$ for X_2.

In systems with more than two variables, one can compute the total information that is exclusively located in one variable as $H_{(1)} \coloneqq \sum_{j=1}^{N} H(X_j|\mathbf{X}_j^c)$, where $\mathbf{X}_j^c$ denotes all of the system's variables, except X_j. The difference between the joint entropy and the sum of all exclusive information terms, $H_{(1)}$, defines a quantity known [25] as the *dual total correlation* (DTC):

$$\mathrm{DTC} = H(\mathbf{X}) - H_{(1)} \ , \tag{7}$$

which measures the portion of the joint entropy that is shared between two or more variables of the system (the superscripts and subscripts are used to reflect that $H^{(1)} \geq H(\mathbf{X}) \geq H_{(1)}$). When $N = 2$, then $\mathrm{DTC} = I(X_1; X_2)$, and hence, the DTC has also been suggested in the literature as a measure for the multivariate mutual information. Note that the DTC is also known as excess entropy in [26], whose definition differs from its typical use of this term in the context of time series, e.g., [27].

By comparing Equations (4) and (7), it would be appealing to look for a decomposition of the DTC of the form $\mathrm{DTC} = \sum_{k=2}^{N} \Delta H_{(k)}$, where $\Delta H_{(k)} \geq 0$ would measure the information that is shared by exactly k variables [28]. With this, one could define an *internal* entropy $H_{(j)} = H_{(1)} + \sum_{i=2}^{j} \Delta H_{(i)}$ as the information that is shared between at most j variables, in contrast to the *external*

entropy $H^{(j)} = H^{(1)} - \sum_{i=2}^{j} \Delta H^{(i)}$, which describes the information provided by the j-marginals. These entropies would form a non-decreasing sequence

$$H_{(1)} \leq \cdots \leq H_{(N-1)} \leq H(\mathbf{X}) \leq H^{(N-1)} \leq \cdots \leq H^{(1)} \ . \tag{8}$$

This layered structure, and its relationship with the TC and the DTC, is graphically represented in Figure 1.

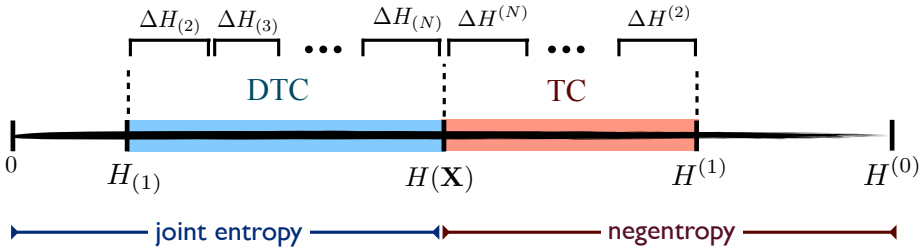

Figure 1. Layers of internal and external entropies that decompose the dual total correlation (DTC) and the TC. Each $\Delta H^{(j)}$ shows how much information is contained in the j-marginals, while each $\Delta H_{(j)}$ measures the information shared between exactly j variables.

One of the main goals of this paper is to find expressions for $\Delta H_{(k)}$ for the case of $N = 3$. It is interesting to note that even though the TC and DTC coincide for the case of $N = 2$, these quantities are in general different for larger system sizes.

2.3. Inclusion-Exclusion Decompositions

Perhaps the most natural approach to decompose the DTC and joint entropy is to apply the inclusion-exclusion principle, using a simplifying analogy that the entropies and areas have similar properties. A refined version of this approach can be found also in the *I-measures* [29] and in the *multi-scale complexity* [30]. For the case of three variables, this approach gives

$$\mathrm{DTC}_{N=3} = I(X_1; X_2|X_3) + I(X_2; X_3|X_1) + I(X_3; X_1|X_2) + I(X_1; X_2; X_3) \ . \tag{9}$$

The last term is known as the *co-information* [31] (being closely related to the *interaction information* [32]) and can be defined using the inclusion-exclusion principle as

$$\begin{aligned} I(X_1; X_2; X_3) :=& H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) \\ & - H(X_1, X_3) + H(X_1, X_2, X_3) \qquad (10) \\ =& I(X_1; X_2) - I(X_1; X_2|X_3) \ . \qquad (11) \end{aligned}$$

As $I(X_1; X_2; X_2) = I(X_1; X_2)$, the co-information has also been proposed as a candidate for extending the mutual information to multiple variables. For a summary of the various possible extensions of the mutual information, see Table 1 and also additional discussion in [33].

Table 1. Summary of the candidates for extending the mutual information.

Name	Formula
Total correlation	$\mathrm{TC} = \sum_j H(X_j) - H(\mathbf{X})$
Dual total correlation	$\mathrm{DTC} = H(\mathbf{X}) - \sum_j H(X_j\|\mathbf{X}_j^c)$
Co-information	$I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2\|X_3)$

It is tempting to coarsen the decomposition provided by Equation (9) and associate the co-information with $\Delta H_{(3)}$ and the remaining terms with $\Delta H_{(2)}$, obtaining a decomposition similar to the one presented in [30]. This idea is equivalent to building a Venn diagram for the information sharing between three variables as the one shown in Figure 2. However, the resulting decomposition and diagram (which differ from what is presented in Section 3.4) are not very intuitive, since the co-information can be negative.

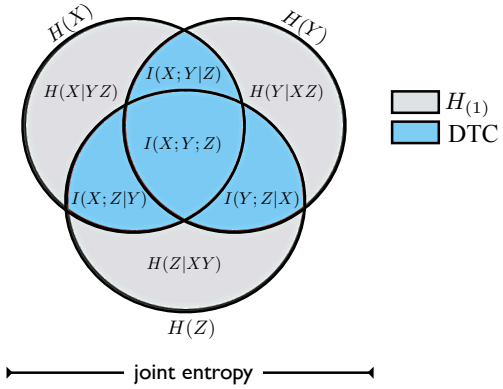

Figure 2. An approach based on the inclusion-exclusion principle decomposes the total entropy of three variables $H(X, Y, Z)$ into seven signed areas.

As part of this temptation, it is appealing to consider the conditional mutual information $I(X_1; X_2|X_3)$ as the information contained in X_1 and X_2 that is not contained in X_3, just as the conditional entropy $H(X_1|X_2)$ is the information that is in X_1 and not in X_2. However, the latter interpretation works because conditioning always reduces entropy (*i.e.*, $H(X_1) \geq H(X_1|X_2)$), while this is not true for mutual information; that is, in some cases, the conditional mutual information $I(X_1; X_2|X_3)$ can be greater than $I(X_1; X_2)$. This suggests that the conditional mutual information can capture information that extends beyond X_1 and X_2, incorporating higher-order effects with respect to X_3. Therefore, a better understanding of the conditional mutual information is required in order to refine the decomposition suggested by Equation (9).

2.4. Synergistic Information

An extended treatment of the conditional mutual information and its relationship to the mutual information decomposition can be found in [34,35]. For presenting these ideas, let us consider two random variables X_1 and X_2, which are used to predict Y. The predictability of Y, understood as the benefit of knowing the realization of X_1 and X_2 for performing inference over Y, is directly related to $I(X_1X_2; Y)$ if a natural data processing property is to be satisfied [36] (for simplicity, through the paper, we use the shorthand notation $X_1X_2 = (X_1, X_2)$). Using the chain rule of the mutual information, the predictability can be decomposed as

$$I(X_1X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \ . \tag{12}$$

It is natural to think that the predictability provided by X_1, which is given by the term $I(X_1; Y)$, can be either *unique* or *redundant* with respect of the information provided by X_2. On the other hand, due to Equation (12), it is clear that the unique predictability contributed by X_2 must be contained in $I(X_2; Y|X_1)$. However, the fact that $I(X_2; Y|X_1)$ can be larger than $I(X_2; Y)$, while the latter contains both the unique and redundant contributions of X_2, suggests that there can be an additional predictability that is accounted for only by the conditional mutual information.

Following this rationale, we denote as *synergistic predictability* the part of the conditional mutual information that corresponds to evidence about the target that is not contained in any single predictor,

but is only revealed when both are known. Is to be noticed that this notion is based on the asymmetrical roles played by predictors and the target. As an example of synergistic predictability, consider again the case in which X_1 and X_2 are independent random bits and $Y = X_1 \texttt{ XOR } X_2$. Then, it can be seen that $I(X_1;Y) = I(X_2;Y) = 0$ but $I(X_1X_2;Y) = I(X_1;Y|X_2) = 1$. Hence, neither X_1 nor X_2 individually provide information about Y, although together, they fully determine it.

Further discussions about the notion of synergistic predictability can be found in [11,12,37–39].

3. A Non-Negative Joint Entropy Decomposition

Following the discussions presented in Sections 2.2 and 2.4, we search for a decomposition of the joint entropy that reflects the private, common and synergistic modes of information sharing. In this way, we want the decomposition to distinguish information that is shared only by a few variables from information that is accessible from the entire system.

Our framework is based on distinguishing the directed notion of *predictability* from the undirected one of *information*. By predictability, we understand "the reduction in optimal prediction risk in the presence of side information" [36]. The predictability is intrinsically a directed notion, which is based on a distinction between predictors and the target variable. On the contrary, we use the term information to exclusively refer to intrinsic statistical properties of the whole system, which do not rely on such a distinction. Currently, there is an ongoing debate about the best way of characterizing and computing predictability—although without using the same terminology—in arbitrary systems (see, for example, [38] and the references therein). Nevertheless, our approach in this work is to explore how far one can reach based on an axiomatic approach. In this way, our results are going to be consistent with any choice of formula that is consistent with the discussed axioms.

In the following, Sections 3.1–3.3 discuss the basic features of predictability and information. After these necessary preliminaries, Section 3.4 finally presents our formal decomposition for the joint entropy of discrete and continuous variables. Note that the rest of this article is focused on the case of three variables, leaving possible extensions for future work.

3.1. Predictability Axioms

Let us consider two variables X_1 and X_2 that are used to predict a target variable $Y := X_3$. Intuitively, $I(X_1;Y)$ quantifies the predictability of Y that is provided by X_1. In the following, we want to find a function $\mathcal{R}(X_1X_2 \to Y)$ that measures the *redundant predictability* provided by X_1 with respect to the predictability provided by X_2, and a function $\mathcal{U}(X_1 \to Y|X_2)$ that measures the *unique predictability* that is provided by X_1, but not by X_2. Following [11], we first determine a number of desired properties that these functions should have.

Definition A *predictability decomposition* is defined by the real-valued functions $\mathcal{R}(X_1X_2 \to Y)$ and $\mathcal{U}(X_1 \to Y|X_2)$ over the distributions of (X_1, Y) and (X_2, Y), which satisfy the following axioms:

(1) Non-negativity: $\mathcal{R}(X_1X_2 \to Y), \mathcal{U}(X_1 \to Y|X_2) \geq 0$.
(2) $I(X_1;Y) = \mathcal{R}(X_1X_2 \to Y) + \mathcal{U}(X_1 \to Y|X_2)$.
(3) $I(X_1X_2;Y) \geq \mathcal{R}(X_1X_2 \to Y) + \mathcal{U}(X_1 \to Y|X_2) + \mathcal{U}(X_2 \to Y|X_1)$.
(4) *Weak symmetry I*: $\mathcal{R}(X_1X_2 \to Y) = \mathcal{R}(X_2X_1 \to Y)$.

The requirement that $\mathcal{R}(X_1X_2 \to Y)$ and $\mathcal{U}(X_1 \to Y|X_2)$ depend only on the pairwise marginals of (X_1, Y) and (X_2, Y), and not on their joint distribution, was first proposed in [39]. Although this property is not required in some of the existent literature (for example, in [38]), most of the predictability decompositions proposed so far respect it [11,37,39]. Furthermore, Axiom (3) states that the sum of the redundant and corresponding unique predictabilities given by each variable cannot be larger than the total predictability —in fact, the difference between the right and left hand terms of Axiom (3) gives the synergistic predictability, whose analysis will not be included in this work to avoid confusing it with the synergistic information, introduced in Section 3.2. Finally, Axiom (4) states that the redundant predictability is independent of the ordering of the predictors.

The following lemma determines the bounds for the redundant predictability (the proof is given in Appendix A).

Lemma 1. *The functions* $\mathcal{R}(X_1X_2 \to Y)$ *and* $\mathcal{U}(X_1 \to Y|X_2) = I(X_1;Y) - \mathcal{R}(X_1X_2 \to Y)$ *satisfy Axioms (1)–(3) if and only if*

$$\min\{I(X_1;Y), I(X_2;Y)\} \geq \mathcal{R}(X_1X_2 \to Y) \geq [I(X_1;X_2;Y)]^+ \ , \tag{13}$$

where $[a]^+ = \max\{a, 0\}$.

Corollary 1. *For variables* X_1, X_2 *and* Y *following an arbitrary joint pdf* $p_{X_1,X_2,Y}$*, there exists at least one predictability decomposition that satisfies Axioms (1)–(4) defined by*

$$\mathcal{R}_{MMI}(X_1X_2 \to Y) := \min\{I(X_1;Y), I(X_2;Y)\} \ . \tag{14}$$

Proof. Being a symmetric function on X_1 and X_2, Equation (14) satisfies Axiom (4). Furthermore, as Equation (14) is equal to the upper bound given in Lemma 1, Axioms (1)–(3) are satisfied due to Lemma 2. □

Above, the subscript MMI corresponds to "minimal mutual information". Note that $\mathcal{R}_{\text{MMI}}$ was introduced in [40] and has also been studied in [12,41].

In principle, the notion of redundant predictability takes the point of view of the target variable and measures the parts that can be predicted by both X_1 and X_2 when they are used by themselves, *i.e.*, without combining them with each other. It is appealing to think that there should exist a unique function that provides such a measure. Nevertheless, these axioms define only very basic properties that a measure of redundant predictability should satisfy, and hence, in general, they are not enough for defining a unique function. In fact, a number of different predictability decompositions have been proposed in the literature [37–39,41].

From all of the candidates that are compatible with the axioms, the decomposition given in Corollary 1 gives the largest possible redundant predictability measure. It is clear that in some cases, this measure gives an over-estimate of the redundant predictability given by X_1 and X_2; for an example of this, consider X_1 and X_2 to be independent variables and $Y = (X_1, X_2)$. Nevertheless, Equation (14) has been proposed as a adequate measure for the redundant predictability of multivariate Gaussians [41] (for a corresponding discussion, see Section 6).

3.2. Shared, Private and Synergistic Information

Let us now introduce an additional axiom, which will form the basis for our proposed *information decomposition*.

Definition A *symmetrical information decomposition* is given by the real valued functions $I_\cap(X_1;X_2;X_3)$ and $I_{\text{priv}}(X_1;X_2|X_3)$ over the marginal distributions of (X_1, X_2), (X_1, X_3) and (X_2, X_3), which satisfy Axioms (1)–(4) for $I_\cap(X_1;X_2;X_3) := \mathcal{R}(X_1X_2 \to X_3)$ and $I_{\text{priv}}(X_1;X_2|X_3) := \mathcal{U}(X_1 \to X_2|X_3)$, while also satisfying the following property:

(5) *Weak symmetry II*: $I_{\text{priv}}(X_1;X_2|X_3) = I_{\text{priv}}(X_2;X_1|X_3)$.

Finally, $I_S(X_1;X_2;X_3)$ is defined as $I_S(X_1;X_2;X_3) := I(X_1;X_2|X_3) - I_{\text{priv}}(X_1;X_2|X_3)$.

The role of Axiom (5) can be related to the role of the fifth of Euclid's postulates, as, while seeming innocuous, their addition has interesting consequences in the corresponding theory. The following lemma explains why this decomposition is denoted as symmetrical and also shows fundamental bounds for these information functions (note that equivalent bounds for predictability decompositions can also be derived using Lemma 1). The proof is presented in Appendix C.

Lemma 2. *The functions that compose a symmetrical information decomposition satisfy the following properties:*

(a) Strong symmetry: $I_\cap(X_1;X_2;X_3)$ and $I_S(X_1;X_2;X_3)$ are symmetric on their three arguments.
(b) Bounds: these quantities satisfy the following inequalities:

$$\begin{aligned}
&\min\{I(X_1;X_2), I(X_2;X_3), I(X_3;X_1)\} \geq I_\cap(X_1;X_2;X_3) \geq [I(X_1;X_2;X_3)]^+, \\
&\min\{I(X_1;X_3), I(X_1;X_3|X_2)\} \geq I_{priv}(X_1;X_3|X_2) \geq 0, \\
&\min\{I(X_1;X_2|X_3), I(X_2;X_3|X_1), I(X_3;X_1|X_2)\} \geq I_S(X_1;X_2;X_3) \geq [-I(X_1;X_2;X_3)]^+.
\end{aligned} \tag{15}$$

The *strong symmetry* property was first presented in [40], where it was shown that it is not compatible with the axioms presented in [11]. Regardless of this, strong symmetry is a highly desirable property when looking for a decomposition of the joint entropy. Discussions about strong symmetry can also be found in [34,38].

A symmetrical information decomposition can be used to decompose the following mutual information as

$$I(X_1X_2;X_3) = I(X_1;X_3) + I(X_2;X_3|X_1) \ , \tag{16}$$
$$I(X_1;X_3) = I_\cap(X_1;X_2;X_3) + I_{\text{priv}}(X_1;X_3|X_2) \ , \tag{17}$$
$$I(X_2;X_3|X_1) = I_{\text{priv}}(X_2;X_3|X_1) + I_S(X_1;X_2;X_3) \ . \tag{18}$$

In contrast to a decomposition based on the predictability, these measures address properties of the system (X_1, X_2, X_3) as a whole, without being dependent on how it is divided between target and predictor variables (for a parallelism with respect to the corresponding predictability measures, see Table 2). Intuitively, $I_\cap(X_1;X_2;X_3)$ measures the *shared information* that is common to X_1, X_2 and X_3; $I_{\text{priv}}(X_1;X_3|X_2)$ quantifies the *private information* that is shared by X_1 and X_3, but not X_2, and $I_S(X_1;X_2;X_3)$ captures the *synergistic information* that exist between (X_1, X_2, X_3). The latter is a non-intuitive mode of information sharing, whose nature we hope to clarify through the analysis of particular cases presented in Sections 4 and 6.

Table 2. Parallelism between predictability and information measures.

Directed Measures	Symmetrical Measures
Redundant predictability $\mathcal{R}(X_1X_2 \to X_3)$	*Shared information* $I_\cap(X_1;X_2;X_3)$
Unique predictability $\mathcal{U}(X_1 \to X_2|X_3)$	*Private information* $I_{\text{priv}}(X_1;X_2|X_3)$
Synergistic predictability	*Synergistic information* $I_S(X_1;X_2;X_3)$

Note also that the co-information can be expressed as

$$I(X_1;X_2;X_3) = I_\cap(X_1;X_2;X_3) - I_S(X_1;X_2;X_3) \ . \tag{19}$$

Hence, strictly positive (resp. negative) co-information is a sufficient, although not necessary, condition for the system to have non-zero shared (resp. synergistic) information.

3.3. Further Properties of the Symmetrical Decomposition

At this point, it is important to clarify a fundamental distinction that we make between the notions of *predictability* and *information*. The main difference between the two notions is that, in principle, the predictability only considers the predictable parts of the target, while the shared information also considers the joint statistics of the predictors. Although this distinction will be

further developed when we address the case of Gaussian variables (*cf.* Section 6.3), let us for now present a simple example to help develop the intuitions about this issue.

Example Define the following functions:

$$I_\cap(X_1;X_2;X_3) = \min\{I(X_1;X_2), I(X_2;X_3), I(X_3;X_1)\} \ , \tag{20}$$

$$I_{\text{priv}}(X_1;X_2|X_3) = I(X_1;X_2) - I_\cap(X_1;X_2;X_3) \ . \tag{21}$$

It is straightforward that these functions satisfy Axioms (1)–(5) and therefore constitute a symmetric information decomposition. In contrast to the decomposition given in Corollary 1, this can be seen to be strongly symmetric and also dependent on the three marginals (X_1, X_2), (X_2, X_3) and (X_1, X_3).

In the following lemma, we will generalize the previous construction, whose simple proof is omitted.

Lemma 3. *For a given predictability decomposition with functions $\mathcal{R}(X_1X_2 \to X_3)$ and $\mathcal{U}(X_1 \to X_2|X_3)$, the functions*

$$I_\cap(X_1;X_2;X_3) = \min\{\mathcal{R}(X_1X_2 \to X_3), \mathcal{R}(X_2X_3 \to X_1), \mathcal{R}(X_3X_1 \to X_2)\} \tag{22}$$

$$I_{priv}(X_1;X_2|X_3) = I(X_1;X_2) - I_\cap(X_1;X_2;X_3) \tag{23}$$

provide a symmetrical information decomposition, which is called the canonical symmetrization of the predictability.

Corollary 2. *For variables X_1, X_2 and X_3 following an arbitrary joint pdf p_{X_1,X_2,X_3}, there exists at least one symmetric information decomposition.*

Proof. This is a direct consequence of the previous lemma and Corollary 1. □

May be the most remarkable property of symmetrized information decompositions is that, in contrast to directed ones, as defined in Section 3.1, they are uniquely determined by Axioms (1)–(5) for a number of interesting cases.

Theorem 1. *The symmetric information decomposition is unique if the variables form a Markov chain or two of them are pairwise independent.*

Proof. Let us consider the upper and lower bound for $I_\cap$ given in Equation (15), denoting them as $c_1 := [I(X_1;X_2;X_3)]^+$ and $c_2 := \min\{I(X_1;X_2), I(X_2;X_3), I(X_1;X_3)\}$. These bounds restrict the possible $I_\cap$ functions to lay in the interval $[c_1, c_2]$ of length

$$|c_2 - c_1| = \min\{I(X_1;X_2), I(X_2;X_3), I(X_1;X_3), I(X_1;X_2|X_3), I(X_2;X_3|X_1), I(X_3;X_1|X_2)\} \ . \tag{24}$$

Therefore, the framework will provide a unique expression for the shared information if (at least) one of the above six terms is zero. These scenarios correspond either to Markov chains, where one conditional mutual information term is zero, or pairwise independent variables, where one mutual information term vanishes. □

Pairwise independent variables and Markov chains are analyzed in Sections 4 and 5.1, respectively.

3.4. Decomposition for the Joint Entropy of Three Variables

Now, we use the notions of redundant, private and synergistic information functions for developing a non-negative decomposition of the joint entropy, which is based on a non-negative

decomposition of the DTC. For the case of three discrete variables, by applying Equations (18) and (19) to Equation (9), one finds that

$$\begin{aligned} \mathrm{DTC} = & I_{\mathrm{priv}}(X_1; X_2|X_3) + I_{\mathrm{priv}}(X_2; X_3|X_1) + I_{\mathrm{priv}}(X_3; X_1|X_2) \\ & + I_{\cap}(X_1; X_2; X_3) + 2I_{\mathrm{S}}(X_1; X_2; X_3) \ . \end{aligned} \tag{25}$$

From Equations (7) and (25), one can propose the following decomposition for the joint entropy:

$$H(X_1, X_2, X_3) = H_{(1)} + \Delta H_{(2)} + \Delta H_{(3)} \ , \tag{26}$$

where

$$H_{(1)} = H(X_1|X_2, X_3) + H(X_2|X_1, X_3) + H(X_3|X_1, X_2) \ , \tag{27}$$

$$\Delta H_{(2)} = I_{\mathrm{priv}}(X_1; X_2|X_3) + I_{\mathrm{priv}}(X_2; X_3|X_1) + I_{\mathrm{priv}}(X_3; X_1|X_2) \ , \tag{28}$$

$$\Delta H_{(3)} = I_{\cap}(X_1; X_2; X_3) + 2I_{\mathrm{S}}(X_1; X_2; X_3) \ . \tag{29}$$

In contrast to Equation (9), here, each term is non-negative because of Lemma 2. Therefore, Equation (26) yields a non-negative decomposition of the joint entropy, where each of the corresponding terms captures the information that is shared by one, two or three variables. Interestingly, $H_{(1)}$ and $\Delta H_{(2)}$ are homogeneous (being the sum of all of the exclusive information or private information of the system), while $\Delta H_{(3)}$ is composed by a mixture of two different information sharing modes. Interestingly, it can be seen from Equations (18) and (19) that the co-information is sometimes negative for compensating the triple counting of the synergy due to the sum of the three conditional mutual information terms.

An analogous decomposition can be developed for the case of continuous random variables. Nevertheless, as the differential entropy can be negative, not all of the terms of the decomposition can be non-negative. In effect, following the same rationale that leads to Equation (26), the following decomposition can be found:

$$h(X_1, X_2, X_3) = h_{(1)} + \Delta H_{(2)} + \Delta H_{(3)} \ . \tag{30}$$

Above, $h(X)$ denotes the differential entropy of X, $\Delta H_{(2)}$ and $\Delta H_{(3)}$ are as defined in Equations (28) and (29), and

$$h_{(1)} = h(X_1|X_2X_3) + h(X_2|X_1X_3) + h(X_3|X_1X_2) \ . \tag{31}$$

Hence, although both the joint entropy $h(X_1, X_2, X_3)$ and $h_{(1)}$ can be negative, the remaining terms conserve their non-negative condition.

It can be seen that the lowest layer of the decomposition is always trivial to compute, and hence, the challenge is to find expressions for $\Delta H_{(2)}$ and $\Delta H_{(3)}$. In the rest of the paper, we will explore scenarios where these quantities can be characterized. Note that in general, $\Delta H_{(k)} \neq \Delta H^{(k)}$, although it is appealing to believe that there should exist a relationship between them. These issues are explored in the following sections.

4. Pairwise Independent Variables

In this section, we focus on the case where two variables are pairwise independent while being globally connected by a third variable. The fact that pairwise independent variables can become correlated when additional information becomes available is known in the statistics literature as the *Bergson's paradox* or *selection bias* [42] or as the *explaining away effect* in the context of artificial intelligence [43]. As an example of this phenomenon, consider X_1 and X_2 to be two pairwise independent canonical Gaussians variables and X_3 a binary variable that is equal to one if $X_1 + X_2 > 0$

and zero otherwise. Then, knowing that $X_3 = 1$ implies that $X_2 > -X_1$, and hence, knowing the value of X_1 effectively reduces the uncertainty about X_2.

In our framework, Bergson's paradox can be understood as synergistic information that is introduced by the third component of the system. In fact, we will show that in this case, the synergistic information function is unique and given by

$$I_S(X_1; X_2; X_3) = \sum_{x_3} p_{X_3}(x_3) I(X_1; X_2 | X_3 = x_3) = I(X_1; X_2 | X_3) \,, \tag{32}$$

which is, in fact, a measure of the dependencies between X_1 and X_2 that are created by X_3. In the following, Section 4.1 presents the unique symmetrized information decomposition for this case. Then, Section 4.2 focuses on the particular case where X_3 is a function of the other two variables.

4.1. Uniqueness of the Entropy Decomposition

Let us assume that X_1 and X_2 are pairwise independent, and hence, the joint pdf of X_1, X_2 and X_3 has the following structure:

$$p_{X_1 X_2 X_3}(x_1, x_2, x_3) = p_{X_1}(x_1) p_{X_2}(x_2) p_{X_3|X_1 X_2}(x_3|x_1, x_2) \ . \tag{33}$$

It is direct to see that in this case $p_{X_1 X_2} = \sum_{x_3} p_{X_1 X_2 X_3} = p_{X_1} p_{X_2}$, but $p_{X_1 X_2|X_3} \neq p_{X_1|X_3} p_{X_2|X_3}$. Therefore, as $I(X_1; X_2) = 0$, it is direct from Axioms (1) and (2) that any redundant predictability function satisfies $\mathcal{R}(X_1 X_3 \to X_2) = \mathcal{R}(X_2 X_3 \to X_1) = 0$. However, the axioms are not enough to uniquely determine $\mathcal{R}(X_1 X_2 \to X_3)$, as the only restriction that the bound presented in Lemma 2 provides is $\min\{I(X_1; X_3), I(X_2; X_3)\} \geq \mathcal{R}(X_1 X_2 \to X_3) \geq 0$ (note that in this case $I(X_1; X_2; X_3) = -I(X_1; X_2 | X_3) \leq 0$). Nevertheless, the symmetrized decomposition is uniquely determined, as shown in the next corollary that is a consequence of Theorem 1.

Corollary 3. *If X_1, X_2 and X_3 follow a pdf as Equation (33), then the shared, private and synergetic information functions are unique. They are given by*

$$I_\cap(X_1; X_2; X_3) = I_{priv}(X_1; X_2 | X_3) = 0 \ , \tag{34}$$

$$I_{priv}(X_1; X_3 | X_2) = I(X_1; X_3) \ , \tag{35}$$

$$I_{priv}(X_2; X_3 | X_1) = I(X_2; X_3) \ , \tag{36}$$

$$I_S(X_1; X_2; X_3) = I(X_1; X_2 | X_3) = -I(X_1; X_2; X_3) \ . \tag{37}$$

Proof. The fact that there is no shared information follows directly from the upper bound presented in Lemma 2. Using this, the expressions for the private information can be found using Axiom (2). Finally, the synergistic information can be computed as

$$I_S(X_1; X_2; X_3) = I(X_1; X_2 | X_3) - I_{\text{priv}}(X_1; X_2 | X_3) = I(X_1; X_2 | X_3) \ . \tag{38}$$

The second formula for the synergistic information can be found then using the fact that $I(X_1; X_2) = 0$. □

With this corollary, the unique decomposition of the DTC $= \Delta H_{(2)} + \Delta H_{(3)}$ that is compatible with Equations (28) and (29) can be found to be

$$\Delta H_{(2)} = I(X_1; X_3) + I(X_2; X_3) \ , \tag{39}$$

$$\Delta H_{(3)} = 2I(X_1; X_2 | X_3) \ . \tag{40}$$

Note that the terms $\Delta H_{(2)}$ and $\Delta H_{(3)}$ can be bounded as follows:

$$\Delta H_{(2)} \leq \min\{H(X_1), H(X_3)\} + \min\{H(X_2), H(X_3)\} \ , \tag{41}$$

$$\Delta H_{(3)} \leq 2\min\{H(X_1|X_3), H(X_2|X_3)\} \ . \tag{42}$$

The bound for $\Delta H_{(2)}$ follows from the basic fact that $I(X;Y) \leq \min\{H(X), H(Y)\}$. The second bound follows from

$$I(X;Y|Z) = \sum_z p_Z(z) I(X;Y|Z=z) \tag{43}$$

$$\leq \sum_z p_Z(z) \min\{H(X|Z=z), H(Y|Z=z)\} \tag{44}$$

$$\leq \min\left\{\sum_z p_Z(z) H(X|Z=z), \sum_z p_Z(z) H(Y|Z=z)\right\} \tag{45}$$

$$= \min\{H(X|Z), H(Y|Z)\} \ . \tag{46}$$

4.2. Functions of Independent Arguments

Let us focus in this section on the special case where $X_3 = F(X_1, X_2)$ is a function of two independent random inputs and study its corresponding entropy decomposition. We will consider X_1 and X_2 as inputs and $F(X_1, X_2)$ to be the output. Although this scenario fits nicely in the predictability framework, it can also be studied from the shared information framework's perspective. Our goal is to understand how F affects the information sharing structure.

As $H(X_3|X_1, X_2) = 0$, we have

$$H_{(1)} = H(X_1|X_2X_3) + H(X_2|X_1X_3) \ . \tag{47}$$

The term $H_{(1)}$ hence measures the information of the inputs that is not reflected by the output. An extreme case is given by a constant function $F(X_1, X_2) = k$, for which $\Delta H_{(2)} = \Delta H_{(3)} = 0$.

The term $\Delta H_{(2)}$ measures how much of F can be predicted with knowledge that comes from one of the inputs, but not from the other. If $\Delta H_{(2)}$ is large, then F is not "mixing" the inputs too much, in the sense that each of them is by itself able to provide relevant information that is not given also by the other. In fact, a maximal value of $\Delta H_{(2)}$ for given marginal distributions for X_1 and X_2 is given by $F(X_1, X_2) = (X_1, X_2)$, where $H_{(1)} = \Delta H_{(3)} = 0$, and the bound provided in Equation (41) is attained.

Finally, due to Equation (34), there is no shared information, and hence, $\Delta H_{(3)}$ is just proportional to the synergy of the system. By considering Equation (42), one finds that F needs to leave some ambiguity about the exact values of the inputs in order for the system to possess synergy. For example, consider a 1-1 function F for which for every output $F(X_1, X_2) = x_3$; one can find the unique values x_1 and x_2 that generate it. Under this condition $H(X_1|X_3) = H(X_2|X_3) = 0$, and hence, because of Equation (42), it is clear that a 1-1 function does not induce synergy. On the other extreme, we showed already that constant functions have $\Delta H_{(3)} = 0$, and hence, the case where the output of the system gives no information about the inputs also leads to no synergy. Therefore, synergistic functions are those whose output values generate a balanced ambiguity about the generating inputs. To develop this idea further, the next lemma studies the functions that generate a maximum amount of synergy by generating for each output value different 1-1 mappings between their arguments.

Lemma 4. *Let us assume that both X_1 and X_2 take values over $\mathcal{K} = \{0, \ldots, K-1\}$ and are independent. Then, the maximal possible amount of information synergy is created by the function*

$$F^*(n, m) = n + m \ (mod \ K) \tag{48}$$

when both input variables are uniformly distributed.

Proof. Using Equations (37) and (46), it can be shown that if F is an arbitrary function, then

$$\begin{aligned} I_{\mathrm{S}}(X_1; X_2; F(X_1, X_2)) &= I(X_1; X_2|F) && (49)\\ &\leq \min\{H(X_1|F), H(X_2|F)\} && (50)\\ &\leq \min\{H(X_1), H(X_2)\} && (51)\\ &\leq \log K \ . && (52) \end{aligned}$$

where the last inequality follows from the fact that both inputs are restricted to alphabets of size K.

Now, consider F^* to be the function given in Equation (48) and assume that X_1 and X_2 are uniformly distributed. It can be seen that for each $z \in \mathcal{K}$, there exist exactly K ordered pairs of inputs (x_1, x_2), such that $F^*(x_1, x_2) = z$, which define a bijection from $\mathcal{K}$ to $\mathcal{K}$. Therefore,

$$I(X_1; X_2|F = z) = H(X_1|z) - H(X_2|X_1, z) = H(X_1) = \log K \tag{53}$$

and hence

$$I_{\mathrm{S}}(X_1; X_2; F^*) = I(X_1; X_2|F^*) = \sum_z \mathbb{P}\{F^* = z\} \cdot I(X_1; X_2|F^* = z) = \log K \ , \tag{54}$$

showing that the upper bound presented in Equation (52) is attained. □

Corollary 4. *The* `XOR` *logic gate generates the largest amount of synergistic information possible for the case of binary inputs.*

The synergistic nature of the addition over finite fields helps to explain the central role it has in various fields. In cryptography, the one-time-pad [44] is an encryption technique that uses finite-field additions for creating a synergistic interdependency between a private message, a public signal and a secret key. This interdependency is completely destroyed when the key is not known, ensuring no information leakage to unintended receivers [45]. Furthermore, in *network coding* [46,47], nodes in the network use linear combinations of their received data packets to create and transmit synergistic combinations of the corresponding information messages. This technique has been shown to achieve multicast capacity in wired communication networks [47] and has also been used to increase the throughput of wireless systems [48].

5. Discrete Pairwise Maximum Entropy Distributions and Markov Chains

This section studies the case where the system's variables follow a *pairwise maximum entropy* (PME) distribution. These distributions are of great importance in the statistical physics and machine learning communities, where they are studied under the names of *Gibbs distributions* [49] or *Markov random fields* [50].

Concretely, let us consider three pairwise marginal distributions $p_{X_1X_2}$, $p_{X_2X_3}$ and $p_{X_1X_3}$ for the discrete variables X_1, X_2 and X_3. Let us denote as $\mathcal{Q}$ the set of all of the joint pdfs over (X_1, X_2, X_3) that have those as their pairwise marginals distributions. Then, the corresponding PME distribution is given by the joint pdf $\tilde{p}_{\mathbf{X}}(x_1, x_2, x_3)$ that satisfies

$$\tilde{p}_{\mathbf{X}} = \underset{p \in \mathcal{Q}}{\operatorname{argmax}}\, H(\{p\}) \ . \tag{55}$$

For the case of binary variables (*i.e.*, $X_j \in \{0, 1\}$), the PME distribution is given by an Ising distribution [51]

$$\tilde{p}_{\mathbf{X}}(\mathbf{X}) = \frac{e^{-\mathcal{E}(\mathbf{X})}}{Z} \ , \tag{56}$$

where Z is a normalization constant and $\mathcal{E}(\mathbf{X})$ an energy function given by $\mathcal{E}(\mathbf{X}) = \sum_i J_i X_i + \sum_j \sum_{k \neq j} J_{j,k} X_j X_k$, being $J_{j,k}$ the coupling terms. In effect, if $J_{i,k} = 0$ for all i and k, then $\tilde{p}_{\mathbf{X}}(\mathbf{X})$ can be factorized as the product of the unary-marginal pdfs.

In the context of the framework discussed in Section 2.1, a PME system has TC $= \Delta H^{(2)}$ while $\Delta H^{(3)} = 0$. In contrast, Section 5.1 studies these systems under the light of the decomposition of the DTC presented in Section 3.4. Then, Section 5.2 specifies the analysis for the particular case of Markov chains.

5.1. Synergy Minimization

It is tempting to associate the synergistic information with that which is only in the joint pdf, but not in the pairwise marginals, *i.e.*, with $\Delta H^{(3)}$. However, the following result states that there can exist some synergy defined by the pairwise marginals themselves.

Theorem 2. *PME distributions have the minimum amount of synergistic information that is allowed by their pairwise marginals.*

Proof. Note that

$$\max_{p \in \mathcal{Q}} H(X_1X_2X_3) = H(X_1X_2) + H(X_3) - \min_{p \in \mathcal{Q}} I(X_1X_2; X_3) \tag{57}$$

$$= H(X_1X_2) + H(X_3) - I(X_1; X_3) - \min_{p \in \mathcal{Q}} I(X_2; X_3|X_1) \tag{58}$$

$$= H(X_1X_2) + H(X_3) - I(X_1; X_3) - I_{\text{priv}}(X_2; X_3|X_1) - \min_{p \in \mathcal{Q}} I_{\text{S}}(X_1; X_2; X_3) \ . \tag{59}$$

Therefore, maximizing the joint entropy for fixed pairwise marginals is equivalent to minimizing the synergistic information. Note that the last equality follows from the fact that by definition $I_{\text{priv}}(X_2; X_3|X_1)$ only depends on the pairwise marginals. □

Corollary 5. *For an arbitrary system (X_1, X_2, X_3), the synergistic information can be decomposed as*

$$I_S(X_1; X_2; X_3) = I_S^{PME} + \Delta H^{(3)} \ , \tag{60}$$

where $\Delta H^{(3)}$ is as defined in Equation (4) and $I_S^{PME} = \min_{p \in \mathcal{Q}} I_S(X_1; X_2; X_3)$ is the synergistic information of the corresponding PME distribution.

Proof. This can be proven noting that, for an arbitrary pdf $p_{X_1X_2X_3}$, it can be seen that

$$\Delta H^{(3)} = \max_{p \in \mathcal{Q}} H(X_1X_2X_3) - H(\{p_{X_1X_2X_3}\}) \tag{61}$$

$$= I_{\text{S}}(\{p_{X_1X_2X_3}\}) - \min_{p \in \mathcal{Q}} I_{\text{S}}(X_1; X_2; X_3) \ . \tag{62}$$

Above, the first equality corresponds to the definition of $\Delta H^{(3)}$, and the second equality comes from using Equation (59) on each joint entropy term and noting that only the synergistic information depends on more than the pairwise marginals. □

The previous corollary shows that $\Delta H^{(3)}$ measures only one part of the information synergy of a system, the part that can be removed without altering the pairwise marginals. Therefore, by considering Equations (29) and (60), one can find that for systems of three variables

$$\Delta H_{(3)} \geq I_{\text{S}}(X_1; X_2; X_3) \geq \Delta H^{(3)} \ . \tag{63}$$

Note that PME systems with non-zero synergy, *i.e.*, systems where the above inequality is strict, are easy to find. For example, consider X_1 and X_2 to be two independent equiprobable bits, and $X_3 = X_1 \ \texttt{AND} \ X_2$. It can be shown that for this case, one has $\Delta H^{(3)} = 0$ [18]. On the other side, as the

inputs are independent, the synergy can be computed using Equation (37), and therefore, a direct calculation shows that

$$I_S(X_1; X_2; X_3) = I(X_1; X_2|X_3) = H(X_1|X_3) - H(X_1|X_2X_3) = 0.1887 \ . \tag{64}$$

From the previous discussion, one can conclude that only a special class of pairwise distributions $p_{X_1X_2}$, $p_{X_1X_3}$ and $p_{X_2X_3}$ is compatible with having null synergistic information. This is a remarkable result, as the synergistic information is usually considered to be an effect purely related to high-order marginals. For example, in [39] property $(**)$ is introduced, which states that for given $p_{X_1X_3}$ and $p_{X_2X_3}$ there exists a joint pdf that is compatible with them while having zero synergistic predictability. In contrast, the above discussion has shown that in our framework, a symmetrized extension of $(**)$ is not true, *i.e.*, it is not always possible to find a joint pdf with zero synergistic information when the three pairwise marginals are given.

It would be interesting to have an expression for the minimal information synergy that a set of pairwise distributions requires, or equivalently, a symmetrized information decomposition for PME distributions. A particular case that allows a unique solution is discussed in the next section.

5.2. Markov Chains

Markov chains maximize the joint entropy subject to constraints on only two of the three pairwise distributions. In effect, following the same rationale as in the proof of Theorem 2, it can be shown that

$$H(X_1, X_2, X_3) = H(X_1X_2) + H(X_3) - I(X_2; X_3) - I(X_1; X_3|X_2) \ . \tag{65}$$

Then, for fixed pairwise distributions $p_{X_1X_2}$ and $p_{X_2X_3}$, maximizing the joint entropy is equivalent to minimizing the conditional mutual information. Moreover, the maximal entropy is attained by the pdf that makes $I(X_1; X_3|X_2) = 0$, which is precisely the Markov chain $X_1 - X_2 - X_3$ with joint distribution

$$p_{X_1X_2X_3} = \frac{p_{X_1X_2}p_{X_2X_3}}{p_{X_2}} \ . \tag{66}$$

For the binary case, it can be shown that a Markov chain corresponds to an Ising distribution, like Equation (56), where the interaction term $J_{1,3}$ is equal to zero.

It is direct to see that two of the redundant predictability terms of a Markov chain $X_1 - X_2 - X_3$ are unique, given by Equation (13) as $\mathcal{R}(X_1X_2 \to X_3) = \mathcal{R}(X_3X_2 \to X_1) = I(X_1; X_2; X_3) = I(X_1; X_3)$. However, the third redundant predictability term, $\mathcal{R}(X_1X_3 \to X_2)$, is not uniquely defined, as the bounds only guarantee

$$\min\{I(X_1; X_2), I(X_2; X_3)\} \geq \mathcal{R}(X_1X_3 \to X_2) \geq I(X_1; X_3) \ . \tag{67}$$

Due to the well-known *data processing inequality* [8], both $I(X_1; X_2)$ and $I(X_2; X_3)$ can be strictly larger than $I(X_1; X_3)$, and hence, this bound does not provide in general a unique definition. In contrast, Theorem 1 showed that the symmetric information decomposition for Markov chains is unique. We develop this decomposition in the following corollary.

Corollary 6. *If $X_1 - X_2 - X_3$ is a Markov chain, then their unique shared, private and synergistic information functions are given by*

$$I_\cap(X_1; X_2; X_3) = I(X_1; X_3) \ , \tag{68}$$

$$I_{priv}(X_1; X_2|X_3) = I(X_1; X_2) - I(X_1; X_3) \ , \tag{69}$$

$$I_{priv}(X_2; X_3|X_1) = I(X_2; X_3) - I(X_1; X_3) \ , \tag{70}$$

$$I_S(X_1; X_2; X_3) = I_{priv}(X_1; X_3|X_2) = 0 \ . \tag{71}$$

In particular, Markov chains have no synergistic information.

Proof. For this case, one can show that

$$\min_{\substack{i,j\in\{1,2,3\}\\ i\neq j}} \{I(X_i;X_j)\} = I(X_1;X_3) = I(X_1;X_2;X_3) \ , \tag{72}$$

where the first equality is a consequence of the data process inequality and the second of the fact that $I(X_1;X_3|X_2) = 0$. The above equality shows that the bounds for the shared information presented in Lemma 2 give the unique solution $I_\cap(X_1;X_2;X_3) = I(X_1;X_3)$. The other equalities are obtained directly using this result and the definition of I_{priv} and I_{S}. □

Using this corollary, the unique decomposition of the DTC $= \Delta H_{(2)} + \Delta H_{(3)}$ for Markov chains that is compatible with Equations (28) and (29) is given by

$$\Delta H_{(2)} = I(X_1;X_2) + I(X_2;X_3) - 2I(X_1;X_3) \ , \tag{73}$$

$$\Delta H_{(3)} = I(X_1;X_3) \ . \tag{74}$$

Hence, Corollary 6 states that a sufficient condition for three pairwise marginals to be compatible with zero information synergy is for them to satisfy the Markov condition $p_{X_3|X_1} = \sum_{X_2} p_{X_3|X_2} p_{X_2|X_1}$. The question of finding a necessary condition is an open problem, intrinsically linked with the problem of finding a good definition for the shared information for arbitrary PME distributions.

To conclude, let us note an interesting duality that exists between Markov chains and the case where two variables are pairwise independent, which is illustrated in Table 3.

Table 3. Duality between Markov chains and pairwise independent variables.

Markov Chains	Pairwise Independent Variables
Conditional pairwise independency	Pairwise independency
$I(X_1;X_3\|X_2) = 0$	$I(X_1;X_2) = 0$
No I_{priv} between X_1 and X_3	No I_{priv} between X_1 and X_2
No synergistic information	No shared information

6. Entropy Decomposition for the Gaussian Case

In this section, we study the entropy-decomposition for the case where (X_1, X_2, X_3) follow a multivariate Gaussian distribution. As the entropy is not affected by translation, we assume, without loss of generality, that all of the variables have zero mean. The covariance matrix is denoted as

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \alpha\sigma_1\sigma_2 & \beta\sigma_1\sigma_3 \\ \alpha\sigma_1\sigma_2 & \sigma_2^2 & \gamma\sigma_2\sigma_3 \\ \beta\sigma_1\sigma_3 & \gamma\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix} \ , \tag{75}$$

where σ_i^2 is the variance of X_i, α is the correlation between X_1 and X_2, β is the correlation between X_1 and X_3 and γ is the correlation between X_2 and X_3. The condition that the matrix Σ should be positive semi-definite yields the following condition

$$1 + 2\alpha\beta\gamma - \alpha^2 - \beta^2 - \gamma^2 \geq 0 \ . \tag{76}$$

Unfortunately, it is not possible to derive directly from Theorem 1 a unique information decomposition for multivariate Gaussian variables with an arbitrary covariance matrix. However, it has been recently shown that the predictability decomposition for multivariate Gaussians is

unique [41]. In the sequel, Sections 6.1 and 6.2 present a discussion on some properties of the shared and synergistic information of multivariate Gaussians, which lead to an intuitive understanding of the predictability decomposition introduced in [41]. Based on this measure, Section 6.3 builds a symmetrical information decomposition and explores some of its consequences.

6.1. Understanding the Synergistic Information Between Gaussians

The simplistic structure of the joint pdf of multivariate Gaussians, which is fully determined by mere second order statistics, could make one think that these systems do not have synergistic information sharing. However, it can be shown that a multivariate Gaussian is the maximum entropy distribution for a given covariance matrix Σ. Hence, the discussion provided in Section 5.1 suggests that these distributions can indeed have non-zero information synergy, depending on the structure of the pairwise distributions or, equivalently, on the properties of Σ.

Moreover, it has been reported that synergistic phenomena are rather common among multivariate Gaussian variables [41]. As a simple example, consider

$$X_1 = A + B \ , \quad X_2 = B \ , \quad X_3 = A \ , \tag{77}$$

where A and B are independent Gaussians. Intuitively, it can be seen that although X_2 is useless by itself for predicting X_3, it can be used jointly with X_1 to remove the noise term B and provide a perfect prediction. For refining this intuition, let us consider a case where the variables have equal variances and X_2 and X_3 are independent (*i.e.*, $\gamma = 0$). Then, the optimal predictor of X_3 given X_1 is $\hat{X}_3^{X_1} = \alpha X_1$; the optimal predictor given X_2 is $\hat{X}_3^{X_2} = 0$; and the optimal predictor given both X_1 and X_2 is [52]

$$\hat{X}_3^{X_1,X_2} = \frac{\beta}{1-\alpha^2}\left(X_1 - \alpha X_2\right) \ . \tag{78}$$

Therefore, although X_2 is useless to predict X_3 by itself, it can be used for further improving the prediction given by X_1. Hence, all of the information provided by X_2 is synergistic, as is useful only when combined with the information provided by X_1. Note that all of these examples fall in the category of the systems considered in Section 4.

6.2. Understanding the Shared Information

Let us start studying the information shared between two Gaussians. For this, let us consider a pair of zero-mean variables (X_1, X_2) with unit variance and correlation α. A suggestive way of expressing these variables is given by

$$X_1 = W_1 \pm W_{12} \ , \quad X_2 = W_2 \pm W_{12} \ , \tag{79}$$

where W_1, W_2 and W_{12} are independent centered Gaussian variables with variances $s_1^2 = s_2^2 = 1 - |\alpha|$ and $s_{12}^2 = |\alpha|$, respectively. Note that the signs in Equation (79) can be set in order to achieve any desired sign for the covariance (as $\mathbb{E}\left\{X_1 X_2\right\} = \pm\mathbb{E}\left\{W_{12}^2\right\} = \pm s_{12}^2$). The mutual information is given by (see Appendix D)

$$I(X_1; X_2) = -(1/2)\log(1-\alpha^2) = -(1/2)\log(1-s_{12}^4) \ , \tag{80}$$

showing that it is directly related to the variance of the common term W_{12}.

For studying the shared information between three Gaussian variables, let us start considering a case where $\alpha = \beta := \rho$ and $\gamma = 0$. It can be seen that (*cf.* Appendix D)

$$I(X_1; X_2; X_3) = \frac{1}{2}\log\frac{1-2\rho^2}{(1-\rho^2)^2} \ . \tag{81}$$

A direct evaluation shows that Equation (81) is non-positive for all ρ with $|\rho| < 1/\sqrt{2}$ (note that $|\rho|$ cannot be larger that $1/\sqrt{2}$ because of Equation (76)). This is consistent with the fact that X_2 and X_3 are pairwise independent, and hence, due to Equation (37), one has that $0 \leq I_S(X_1; X_2; X_3) = -I(X_1; X_2; X_3)$. Therefore, from Equations (19) and (34) is clear that this system has no shared information for all ρ and has zero synergistic information only for $\rho = 0$.

In contrast, let us now consider a case where $\alpha = \beta = \gamma := \rho > 0$, for which

$$I(X_1; X_2; X_3) = \frac{1}{2} \log \frac{1 + 2\rho^3 - 3\rho^2}{(1-\rho^2)^3} \quad . \tag{82}$$

A direct evaluation shows that, in contrast to Equation (81), the co-information in this case is non-negative, showing that the system is dominated by shared information for all $\rho \neq 0$.

The previous discussion suggests that the shared information depends on the smallest of the correlation coefficients. An interesting approach to understand this fact can be found in [41], where the predictability among Gaussians is discussed. In this work, the author note that from the point of view of X_3, both X_1 and X_2 are able to decompose the target in a predictable and an unpredictable portion: $X_3 = \hat{X}_3 + E$. In this sense, both predictors achieve the same effect, although with a different efficiency, which is determined by their correlation coefficient. As a consequence of this, the predictor that is less correlated with the target does not provide unique predictability, and hence, its contribution is entirely redundant. This motivates using $\mathcal{R}_{\text{MMI}}(X_1 X_2 \to X_3)$, as defined in Equation (14), as an adequate measure for the redundant predictability among Gaussians. Moreover, in [41], it is shown that $\mathcal{R}_{\text{MMI}}(X_1 X_2 \to X_3)$ is the only consistent definition of redundant predictability that only depends on the pairwise distributions of (X_1, X_3) and (X_2, X_3).

6.3. Shared, Private and Synergistic Information for Gaussian Variables

Based on the previous discussion, the unique definition of shared information among Gaussians that corresponds to a canonical symmetrization of a redundant predictability measure (as discussed in Lemma 3) is given by

$$\begin{aligned} I_\cap(X_1; X_2; X_3) &= \min\{\mathcal{R}_{\text{MMI}}(X_1 X_2 \to X_3), \mathcal{R}_{\text{MMI}}(X_2 X_3 \to X_1). \mathcal{R}_{\text{MMI}}(X_3 X_1 \to X_2)\} && (83) \\ &= \min\{I(X_1; X_2), I(X_2; X_3), I(X_1; X_3)\} && (84) \\ &= -\frac{1}{2} \log(1 - \min\{\alpha^2, \beta^2, \gamma^2\}) \ . && (85) \end{aligned}$$

In contrast with $\mathcal{R}_{\text{MMI}}(X_1 X_2 \to X_3)$, Equation (85) states that there cannot be information shared by the three components of the system if two of them are pairwise independent. Therefore, the magnitude of the shared information is governed by the lowest correlation coefficient of the whole system, being upper-bounded by any of the redundant predictability terms. An intuitive understanding of this definition can be built over a subclass of multivariate Gaussians using the following lemma (whose proof is presented in Appendix E).

Lemma 5. *Let (X_1, X_2, X_3) follow a multivariate Gaussian distribution with zero mean and covariance matrix Σ, as given in Equation (75). Let us further assume that $\alpha \geq \beta \geq \gamma \geq 0$ and $1 - \alpha - \beta + \gamma \geq 0$. Then,*

$$\begin{aligned} \frac{X_1}{\sigma_1} &= s_{123} W_{123} + s_{12} W_{12} + s_{13} W_{13} + s_1 W_1 && (86) \\ \frac{X_2}{\sigma_2} &= s_{123} W_{123} + s_{12} W_{12} + s_2 W_2 && (87) \\ \frac{X_3}{\sigma_3} &= s_{123} W_{123} + s_{13} W_{13} + s_3 W_3 && (88) \end{aligned}$$

where $W_{123}, W_{12}, W_{13}, W_1, W_2$ *and* W_3 *are independent standard Gaussians and* $s_{123}, s_{12}, s_{13}, s_1, s_2$ *and* s_3 *are given by*

$$\begin{aligned} s_{123} &= \sqrt{\gamma}, \quad s_{12} = \sqrt{\alpha - \gamma}, \quad s_{13} = \sqrt{\beta - \gamma}, \\ s_1 &= \sqrt{1 - \alpha - \beta + \gamma}, \quad s_2 = \sqrt{1 - \alpha}, \quad s_3 = \sqrt{1 - \beta}. \end{aligned} \tag{89}$$

It is natural to relate s_{123} to the shared information, s_{12} and s_{13} to the private information and s_1, s_2 and s_3 to the exclusive terms. Moreover, if $\alpha \geq \beta \geq \gamma \geq 0$ do not hold, then $s_{123}^2 = \min\{|\alpha|, |\beta|, |\gamma|\}$, which is consistent with what Equations (79) and (80) state for the case of two variables. Note that the coefficients determined in Equation (89) are unique, as they correspond to the six degrees of freedom of the variables (X_1, X_2, X_3). Finally, note also that the decomposition presented in Lemma 5 does not require a private component between the least correlated variables, *i.e.*, a term W_{23}.

Based on Equation (85), the remaining elements of a symmetric information decomposition for Gaussians can be found as

$$I_{\text{priv}}(X_1; X_2 | X_3) = I(X_1; X_2) - I_\cap(X_1; X_2; X_3) \tag{90}$$

$$= \frac{1}{2} \log \frac{1 - \min\{\alpha^2, \beta^2, \gamma^2\}}{1 - \alpha^2} \ , \tag{91}$$

$$I_{\text{S}}(X_1; X_2; X_3) = I(X_1; X_2 | X_3) - I_{\text{priv}}(X_1; X_2 | X_3) \tag{92}$$

$$= \frac{1}{2} \log \frac{(1 - \alpha^2)(1 - \beta^2)(1 - \gamma^2)}{(1 + 2\alpha\beta\gamma - \alpha^2 - \beta^2 - \gamma^2)(1 - \min\{\alpha^2, \beta^2, \gamma^2\})} \ . \tag{93}$$

According to Equation (91), the two less correlated Gaussians share no private information, which is consistent with the absence of W_{23} in Lemma 5. Moreover, by comparing Equations (93) and (D5), it can be seen that if X_1 and X_2 are the less correlated variables, then the synergistic information can be expressed as $I_{\text{S}}(X_1; X_2; X_3) = I(X_1; X_2 | X_3)$, which, for the particular case of $\alpha = 0$, confirms Equation (37). This, in turn, also shows that, for the particular case of Gaussians variables, forming a Markov chain is a necessary and sufficient condition for having zero information synergy (for the case of $\alpha \geq \beta \geq \gamma$, a direct calculation shows that $I(X_1; X_2 | X_3) = 0$ is equivalent to $\gamma = \alpha\beta$).

To close this section, let us note that Equation (84) corresponds to the upper bound provided by Equation (15), which means that multivariate Gaussians have a maximal shared information. This is complementary to the fact that, because of being a maximum entropy distribution, they also have the smallest amount of synergy that is compatible with the corresponding second order statistics.

7. Applications to Network Information Theory

In this section, we use the framework presented in Section 3 to analyze four fundamental scenarios in network information theory [53]. Our goal is to illustrate how the framework can be used to build new intuitions over these well-known optimal information-theoretic strategies. The application of the framework to scenarios with open problems is left for future work.

In the following, Section 7.1 uses the general framework to analyze the Slepian–Wolf coding for three sources, which is a fundamental result in the literature of distributed source compression. Then, Section 7.2 applies the results of Section 4 to the multiple access channel, which is one of the fundamental settings in multiuser information theory. Section 7.3 uses the results related to Markov chains from Section 5 to the wiretap channel, which constitutes one of the main models of information-theoretic secrecy. Finally, Section 7.4 uses results from Section 6 to study fundamental limits of public or private broadcast transmissions over Gaussian channels.

7.1. Slepian–Wolf Coding

The Slepian–Wolf coding gives lower bounds for the data rates that are required in order to transfer the information contained in various data sources. Let us denote as R_k the data rate of the

k-th source and define $\tilde{R}_k = R_k - H(X_k|\mathbf{X}_k^c)$ as the extra data rate that each source has above its own exclusive information (*cf.* Section 2.2). Then, in the case of two sources X_1 and X_2, the well-known Slepian–Wolf bounds can be re-written as $\tilde{R}_1 \geq 0$, $\tilde{R}_2 \geq 0$, and $\tilde{R}_1 + \tilde{R}_2 \geq I(X_1; X_2)$ ([53], Section 10.3). The last inequality states that $I(X_1; X_2)$ corresponds to shared information that can be transmitted by any of the two sources.

Let us consider now the case of three sources, and denote $R_S = I_S(X_1; X_2; X_3)$. The Slepian–Wolf bounds provide seven inequalities ([53], Section 10.5), which can be re-written as

$$\tilde{R}_i \geq 0, \; i \in \{1,2,3\} \;, \tag{94}$$

$$\tilde{R}_i + \tilde{R}_j \geq I_{\text{priv}}(X_i; X_j|X_k) + R_S \;\; \text{for } i,j,k \in \{1,2,3\}, \; i < j \;, \tag{95}$$

$$\tilde{R}_1 + \tilde{R}_2 + \tilde{R}_3 \geq \Delta H_{(2)} + \Delta H_{(3)} \;. \tag{96}$$

Above, Equation (96) states that the DTC needs to be accounted for by the extra rate of the sources and Equation (95) that every pair needs to take care of its private information. Interestingly, due to Equation (29), the shared information needs to be included in only one of the rates, while the synergistic information needs to be included in at least two. For example, one possible solution that is consistent with these bounds is $\tilde{R}_1 = I_\cap(X_1; X_2; X_3) + I_{\text{priv}}(X_1; X_2|X_3) + I_{\text{priv}}(X_1; X_3|X_3) + I_S(X_1; X_2; X_3)$, $\tilde{R}_2 = I_{\text{priv}}(X_2; X_3|X_1) + I_S(X_1; X_2; X_3)$ and $\tilde{R}_3 = 0$. This can be interpreted as follows: it is sufficient if just one of the three sources provides the information $I_\cap$ shared among all sources and if for each pair of sources, exactly one transfers their private information I_{priv}. Additionally, the constraints on synergistic information require that exactly two of the three sources transfer it. This brings some light to the factor of two required in Equation (29).

7.2. Multiple Access Channel

Let us consider a multiple access channel, where two pairwise independent transmitters send X_1 and X_2 and a receiver gets X_3, as shown in Figure 3. It is well known that, for a given distribution $(X_1, X_2) \sim p(x_1)p(x_2)$, the achievable transmission rates R_1 and R_2 satisfy the constrains given by ([53], Section 4.5)

$$R_1 \leq I(X_1; X_3|X_2) \;, \quad R_2 \leq I(X_2; X_3|X_1) \;, \quad R_1 + R_2 \leq I(X_1, X_2; X_3) \;. \tag{97}$$

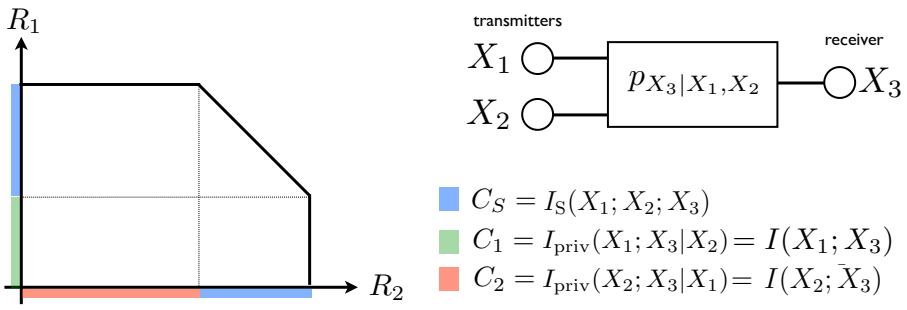

Figure 3. Capacity region of the multiple access channel, which represents the possible data rates that two transmitters can use for transferring information to one receiver.

As the transmitted random variables are pairwise independent, one can apply the results of Section 4. Therefore, there is no shared information, and $I_S(X_1; X_2; X_3) = I(X_1; X_3|X_2) - I(X_1; X_3)$. Let us introduce a shorthand notation for the remaining terms: $C_1 = I_{\text{priv}}(X_1; X_3|X_2) = I(X_1; X_3)$,

$C_2 = I_{\text{priv}}(X_2; X_3|X_1) = I(X_2; X_3)$ and $C_S = I_S(X_1; X_2; X_3)$. Then, one can re-write the bounds for the transmission rates as

$$R_1 \leq C_1 + C_S \ , \quad R_2 \leq C_2 + C_S \ , \quad R_1 + R_2 \leq C_1 + C_2 + C_S \ . \tag{98}$$

From this, it is clear that while each transmitter has a private portion of the channel with capacity C_1 or C_2, their interaction creates *synergistically* extra capacity C_S that corresponds to what can be actually shared.

7.3. Degraded Wiretap Channel

Consider a communication system with an eavesdropper (shown in Figure 4), where the transmitter sends X_1, the intended receiver gets X_2 and the eavesdropper receives X_3. For simplicity of the exposition, let us consider the case where the eavesdropper gets only a degraded copy of the signal received by the intended receiver, *i.e.*, that $X_1 - X_2 - X_3$ form a Markov chain. Using the results of Section 5.2, one can see that in this case, there is no synergistic, but only shared and private information between X_1, X_2 and X_3.

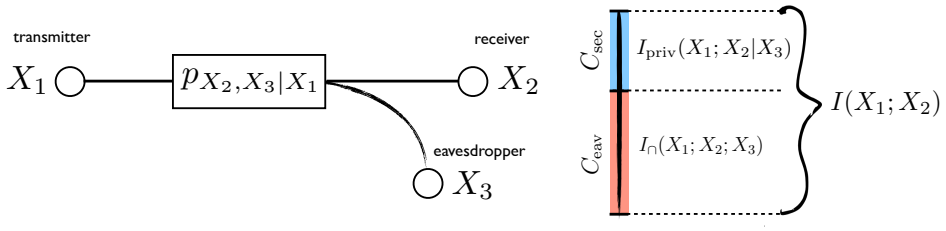

Figure 4. The rate of secure information transfer, C_{sec}, is the portion of the mutual information that can be used while providing perfect confidentiality with respect to the eavesdropper.

In this scenario, it is known that for a given input distribution p_{X_1}, the rate of secure communication that can be achieved is upper bounded by ([44], Section 3.4)

$$C_{\text{sec}} = I(X_1; X_2) - I(X_1; X_3) = I_{\text{priv}}(X_1; X_2|X_3) \ , \tag{99}$$

which is precisely the private information sharing between X_1 and X_2. Furthermore, as intuition would suggest, the eavesdropping capacity is equal to the shared information between the three variables:

$$C_{\text{eav}} = I(X_1; X_2) - C_{\text{sec}} = I(X_1; X_3) = I_\cap(X_1; X_2; X_3) \ . \tag{100}$$

7.4. Gaussian Broadcast Channel

Let us consider a Gaussian broadcast channel, where a transmitter sends a Gaussian signal X_1 that is received as X_2 and X_3 by two receivers. Assuming that all of these variables are jointly Gaussian with a zero mean and covariance matrix given by Equation (75), the transmitter can broadcast a public message, intended for both users, at a maximum rate C_{pub} given by ([44], Section 5.1)

$$C_{\text{pub}} = \min\{I(X_1; X_2), I(X_1; X_3)\} = \mathcal{R}_{\text{MMI}}(X_2 X_3 \to X_1) \ , \tag{101}$$

where, following the discussion presented in Section 6.2, the MMI redundant predictability $\mathcal{R}_{\text{MMI}}(X_2 X_3 \to X_1)$ is as defined in Equation (14). On the other hand, if the transmitter wants to send a private (confidential) message to Receiver 1, the corresponding maximum rate C_{priv} that can be achieved in this case is given by

$$C_{\text{priv}} = [I(X_1; X_2) - I(X_1; X_3)]^+ = I(X_1; X_2) - \mathcal{R}_{\text{MMI}}(X_2 X_3 \to X_1) = \mathcal{U}(X_1 \to X_2|X_3) \ , \tag{102}$$

where the last equality follows from Axiom (2).

Interestingly, the predictability measures prove to be better suited to describing the communication limits in the above scenario than their symmetrical counterparts. In effect, the shared information, as defined in Section 6.3, underestimates the public capacity. This opens the question of whether or not directed measures could be better suited for studying certain communication systems, compared to their symmetrized counterparts. Unfortunately, one cannot explore this issue in the previous cases, as although the symmetric decomposition is uniquely defined, there is no unique predictability decomposition to compare. Therefore, a definite answer to this question is not straightforward at this stage. We hope that future research will provide more evidence and a better understanding of this issue.

8. Conclusions

In this work, we propose an axiomatic framework for studying the interdependencies that can exist between multiple random variables as different modes of information sharing. The framework is based on a symmetric notion of information that refers to properties of the system as a whole. We showed that, in contrast to predictability-based decompositions, all of the information terms of the proposed decomposition have unique expressions for Markov chains and for the case where two variables are pairwise independent. We also analyzed the cases of pairwise maximum entropy (PME) distributions and multivariate Gaussian variables. Finally, we illustrated the application of the framework by using it to develop a more intuitive understanding of the optimal information-theoretic strategies in several fundamental communication scenarios. These results are focused on the case of three variables, as their generalization to a larger number of variables is not straightforward. One of the main difficulties for such a generalization is the increasing complexity of the decompositions, caused by the exponential growth of the number of information-sharing modes [11]).

The key insight that this framework provides is that although there is only one way in which information can be shared between two random variables, there are two essentially different ways of sharing between three. One of these ways is a simple extension of the pairwise dependency, where information is shared redundantly, and hence, any of the variables can be used to predict any other. The second way leads to the counter-intuitive notion of synergistic information sharing, where the information is shared in a way that the statistical dependency is destroyed if any of the variables is removed; hence, the structure exists in the whole, but not in any of the parts.

The synergistic information has been commonly related to statistical structures that exist only in the joint pdf and not in low-order marginals. Interestingly, although we showed that indeed, PME distributions posses the minimal information synergy that is allowed by their pairwise marginals, this minimum can be strictly positive. Therefore, there exists a connection between pairwise marginals and synergistic information sharing that is still to be further clarified. In fact, this phenomenon is related to the difference between the TC and the DTC, which is rooted in the fact that the information sharing modes and the marginal structure of the pdf are, although somehow related, intrinsically different. This important distinction has been represented in our framework by the sequence of internal and external entropies. This new unifying picture for the entropy, negentropy, TC and DTC has shed new light on the understanding of high-order interdependencies, whose consequences have only begun to be explored.

Acknowledgments: We want to thank David Krakauer and Jessica Flack for providing support and inspiration for this research. We also thank Bryan Daniels, Michael Gastpar, Bernhard Geiger, Vigil Griffith and Martin Ugarte for helpful discussions. This work was partially supported by a grant to the Santa Fe Institute for the study of complexity and by the U.S. Army Research Laboratory and the U.S. Army Research Office under Contract Number W911NF-13-1-0340. Fernando Rosas would also like to acknowledge the support of the F+ fellowship from KU Leuven and the project "SINS: Sound INterfacing through the Swarm", funded by the Agency for Innovation by Science and Technology (IWT), Belgium.

Author Contributions: The research was initiated by Fernando Rosas and carried out by all of the authors. The manuscript was initially written by Fernando Rosas and then edited by the other authors. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix

A. Proof of Lemma 1

Proof. Let us assume that $\mathcal{R}(X_1X_2 \to Y)$ and $\mathcal{U}(X_1 \to Y|X_2) = I(X_1;Y) - \mathcal{R}(X_1X_2 \to Y)$ satisfy Axioms (1)–(3). Then,

$$\begin{aligned} I(X_1;Y) &\geq I(X_1;Y) - \mathcal{U}(X_1 \to Y|X_2) && \text{(A1)} \\ &= \mathcal{R}(X_1X_2 \to Y) && \text{(A2)} \\ &= I(X_2;Y) - \mathcal{U}(X_2 \to Y|X_1) \leq I(X_2;Y) \ , && \text{(A3)} \end{aligned}$$

where the inequalities are a consequence of the non-negativity of $\mathcal{U}(X_1 \to Y|X_2)$ and the third equality is due to the weak symmetry of the redundant predictability. To verify the lower bound, first notice that Axiom (3) can be re-written as

$$I(X_1X_2;Y) \geq I(X_1;Y) + I(X_2;Y) - \mathcal{R}(X_1X_2 \to Y). \tag{A4}$$

The lower bound follows considering the non-negativity of $\mathcal{R}(X_1X_2 \to Y)$ and by noting that $I(X_1;Y) + I(X_2;Y) - I(X_1X_2;Y) = I(X_1;X_2;Y)$.

The proof of the converse is direct and left as an exercise to the reader. □

B. Proof of the Consistency of Axiom (3)

Let us show that $\min\{I(X_1;X_2), I(X_1;X_2)\} \geq I(X_1;X_2;X_3)$, showing that the bounds defined by Axiom (3) always can be satisfied. For this, let us assume that the variables are ordered in a way such that $I(X_1;X_2) = \min\{I(X_1;X_2), I(X_2;X_3), I(X_3;X_1)\}$ holds. Then, as one can express $I(X_1;X_2;X_3) = I(X_1,X_2) - I(X_1,X_2|X_3)$, it is direct to show that

$$\begin{aligned} \min\{I(X_1;X_2), I(X_1;X_2)\} - I(X_1;X_2;X_3) &\geq I(X_1;X_2) - I(X_1;X_2;X_3) && \text{(B1)} \\ &= I(X_1;X_2|X_3) && \text{(B2)} \\ &\geq 0 \ , && \text{(B3)} \end{aligned}$$

from where the desired result follows.

C. Proof of Lemma 2

Proof. The symmetry of $I_\cap(X_1;X_2;X_3)$ with respect to X_1 and X_2 can be directly verified from its definition and Axiom (4). The symmetry with respect to X_1 and X_3 is proved using Axiom (2) and the weak symmetry of $I_{\text{priv}}(X_1;X_3|X_2)$ as follows:

$$\begin{aligned} I_\cap(X_1;X_2;X_3) &= I(X_1;X_3) - I_{\text{priv}}(X_1;X_3|X_2) && \text{(C1)} \\ &= I(X_3;X_1) - I_{\text{priv}}(X_3;X_1|X_2) && \text{(C2)} \\ &= I_\cap(X_3;X_2;X_1) \ . && \text{(C3)} \end{aligned}$$

The strong symmetry of $I_S(X_1;X_2;X_3)$ is proved directly using (19) and the strong symmetry of the shared information and the co-information.

The bounds for $I_\cap(X_1;X_2;X_3)$, $I_{\text{priv}}(X_1;X_2;X_3)$ and $I_S(X_1;X_2;X_3)$ follow directly from the definition of these quantities and Axiom (3). □

D. Useful Facts about Gaussians

Here, we list some useful expressions for Gaussian variables:

$$I(X_1;X_2) = \frac{1}{2}\log\frac{1}{1-\alpha^2} \tag{D1}$$

$$= \frac{1}{2}\log\frac{\sigma^2}{|\Sigma_{12}|} \ , \tag{D2}$$

$$I(X_1;X_2,X_3) = \frac{1}{2}\log\frac{1-\gamma^2}{1+2\alpha\beta\gamma-\alpha^2-\beta^2-\gamma^2} \tag{D3}$$

$$= \frac{1}{2}\log\frac{|\Sigma_{23}|}{|\Sigma|} \ , \tag{D4}$$

$$I(X_1;X_2|X_3) = \frac{1}{2}\log\frac{(1-\beta^2)(1-\gamma^2)}{1+2\alpha\beta\gamma-\alpha^2-\beta^2-\gamma^2} \tag{D5}$$

$$= \frac{1}{2}\log\frac{|\Sigma_{13}\Sigma_{23}|}{|\Sigma|} \ , \tag{D6}$$

$$I(X_1;X_2;X_3) = \frac{1}{2}\log\frac{1+2\alpha\beta\gamma-\alpha^2-\beta^2-\gamma^2}{(1-\alpha^2)(1-\beta^2)(1-\gamma^2)} \tag{D7}$$

$$= \frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_{12}\Sigma_{13}\Sigma_{23}|} \ , \tag{D8}$$

where $|\Delta|$ is a matrix determinant, and

$$\Sigma_{12} = \begin{pmatrix} \sigma^2 & \alpha\sigma^2 \\ \alpha\sigma^2 & \sigma^2 \end{pmatrix} \qquad \Sigma_{13} = \begin{pmatrix} \sigma^2 & \beta\sigma^2 \\ \beta\sigma^2 & \sigma^2 \end{pmatrix} \qquad \Sigma_{23} = \begin{pmatrix} \sigma^2 & \gamma\sigma^2 \\ \gamma\sigma^2 & \sigma^2 \end{pmatrix} \ . \tag{D9}$$

E. Proof of Lemma 5

Proof. Consider the following random variables:

$$Y_1 = \sigma_1(s_{123}W_{123} + s_{12}W_{12} + s_{13}W_{13} + s_1W_1) \ , \tag{E1}$$

$$Y_2 = \sigma_2(s_{123}W_{123} + s_{12}W_{12} + s_2W_2) \ , \tag{E2}$$

$$Y_3 = \sigma_3(s_{123}W_{123} + s_{13}W_{13} + s_3W_3) \ , \tag{E3}$$

where $W_{123}, W_{12}, W_{13}, W_1, W_2$ and W_3 are independent standard Gaussians and the parameters $s_{123}, s_{12}, s_{13}, s_1, s_2$ and s_3 as defined in (89). Then, it is direct to check that $\mathbf{Y} = (Y_1, Y_2, Y_3)$ is a multivariate Gaussian variable with zero mean and covariance matrix $\Sigma_{\mathbf{Y}}$ equal to (75). Therefore, (Y_1, Y_2, Y_3) and (X_1, X_2, X_3) have the same statistics, which proves the desired result. □

References

1. Kaneko, K. *Life: An Introduction to Complex Systems Biology*; Springer-Verlag: Berlin/Heidelberg, Germany, 2006.
2. Perrings, C. *Economy and Environment: A Theoretical Essay on the Interdependence of Economic and Environmental Systems*; Cambridge University Press: Cambridge, UK, 2005.
3. Martignon, L.; Deco, G.; Laskey, K.; Diamond, M.; Freiwald, W.; Vaadia, E. Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Comput.* **2000**, *12*, 2621–2653.
4. Deutscher, D.; Meilijson, I.; Schuster, S.; Ruppin, E. Can single knockouts accurately single out gene functions? *BMC Syst. Biol.* **2008**, *2*, 50, doi:10.1186/1752-0509-2-50.
5. Anand, K.; Bianconi, G. Entropy measures for networks: Toward an information theory of complex topologies. *Phys. Rev. E* **2009**, *80*, 045102.

6. Gastpar, M.; Vetterli, M.; Dragotti, P.L. Sensing reality and communicating bits: A dangerous liaison. *IEEE Signal Process. Mag.* **2006**, *23*, 70–83.
7. Casella, G.; Berger, R.L. *Statistical Inference*; Duxbury Press: Pacific Grove, CA, USA, 2002.
8. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley: Hoboken, NJ, USA, 1991.
9. Senge, P.M.; Smith, B.; Kruschwitz, N.; Laur, J.; Schley, S. *The Necessary Revolution: How Individuals and Organizations Are Working Together to Create a Sustainable World*; Crown Business: New York, NY, USA, 2008.
10. Amari, S.I. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711.
11. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. **2010**, arXiv:1004.2515.
12. Olbrich, E.; Bertschinger, N.; Rauh, J. Information Decomposition and Synergy. *Entropy* **2015**, *17*, 3501–3517.
13. Li, W. Mutual information functions versus correlation functions. *J. Stat. Phys.* **1990**, *60*, 823–837.
14. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
15. Brillouin, L. The negentropy principle of information. *J. Appl. Phys.* **1953**, *24*, 1152–1163.
16. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82.
17. Studenỳ, M.; Vejnarová, J. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in Graphical Models*; Springer Netherlands: Amsterdam, The Netherlands, 1998; pp. 261–297.
18. Schneidman, E.; Still, S.; Berry, M.J.; Bialek, W. Network information and connected correlations. *Phys. Rev. Lett.* **2003**, *91*, 238701.
19. Schneidman, E.; Berry, M.J.; Segev, R.; Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012.
20. Roudi, Y.; Nirenberg, S.; Latham, P.E. Pairwise Maximum Entropy Models for Studying Large Biological Systems: When They Can Work and When They Can't. *PLoS Comput. Biol.* **2009**, *5*, e1000380.
21. Bialek, W.; Cavagna, A.; Giardina, I.; Mora, T.; Silvestri, E.; Viale, M.; Walczak, A.M. Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 4786–4791.
22. Merchan, L.; Nemenman, I. On the sufficiency of pairwise interactions in maximum entropy models of biological networks. **2015**, arXiv:1505.02831.
23. Daniels, B.C.; Krakauer, D.C.; Flack, J.C. Sparse code of conflict in a primate society. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 14259–14264.
24. Lee, E.D.; Broedersz, C.P.; Bialek, W. Statistical mechanics of the US Supreme Court. *J. Stat. Phys.* **2013**, *160*, 275–301.
25. Sun, H.T. Nonnegative entropy measures of multivariate symmetric correlations. *Inf. Control* **1978**, *36*, 133–156.
26. Olbrich, E.; Bertschinger, N.; Ay, N.; Jost, J. How should complexity scale with system size? *Eur. Phys. J. B* **2008**, *63*, 407–415.
27. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos* **2003**, *13*, 25–54.
28. Rosas, F.; Ntranos, V.; Ellison, C.J.; Verhelst, M.; Pollin, S. Understanding high-order correlations using a synergy-based decomposition of the total entropy. In Proceedings of the 5th Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux, Brussels, Belgium, 6–7 May 2015; pp. 146–153.
29. Yeung, R.W. A new outlook on Shannon's information measures. *IEEE Trans. Inf. Theory* **1991**, *37*, 466–474.
30. Bar-Yam, Y. Multiscale complexity/entropy. *Adv. Complex Syst.* **2004**, *7*, 47–63.
31. Bell, A.J. The co-information lattice. In Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003), Nara, Japan, 1–4 April 2003; pp. 921–926.
32. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
33. James, R.G.; Ellison, C.J.; Crutchfield, J.P. Anatomy of a bit: Information in a time series observation. *Chaos Interdiscip. J Nonlinear Sci.* **2011**, *21*, 037109.
34. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer-Verlag: Berlin/Heidelberg, Germany, 2014; Volume 9; pp. 159–190.
35. Griffith, V. Quantifying synergistic information. Ph.D. Thesis, California Institute of Technology, Pasadena, CA, USA, 2014.
36. Jiao, J.; Courtade, T.; Venkat, K.; Weissman, T. Justification of Logarithmic Loss via the Benefit of Side Information. *IEEE Trans. Inf. Theory* **2015**, *61*, 5357–5365.
37. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130.

38. Griffith, V.; Chong, E.K.; James, R.G.; Ellison, C.J.; Crutchfield, J.P. Intersection information based on common randomness. *Entropy* **2014**, *16*, 1985–2000.
39. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183.
40. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared information—New insights and problems in decomposing information in complex systems. In Proceedings of the European Conference on Complex Systems 2012, Brussels, Belgium, 3–7 September 2012; pp. 251–269.
41. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E* **2015**, *91*, 052802.
42. Berkson, J. Limitations of the application of fourfold table analysis to hospital data. *Biom. Bull.* **1946**, *2*, 47–53.
43. Kim, J.; Pearl, J. A computational model for causal and diagnostic reasoning in inference systems. In Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI), Karlsruhe, Germany, 8–12 August 1983; pp. 190–193.
44. Bloch, M.; Barros, J. *Physical-Layer Security: From Information Theory to Security Engineering*; Cambridge University Press: Cambrige, UK, 2011.
45. Shannon, C.E. Communication theory of secrecy systems*. *Bell Syst. Tech. J.* **1949**, *28*, 656–715.
46. Ahlswede, R.; Cai, N.; Li, S.Y.R.; Yeung, R.W. Network information flow. *IEEE Trans. Inf. Theory* **2000**, *46*, 1204–1216.
47. Li, S.Y.R.; Yeung, R.W.; Cai, N. Linear network coding. *IEEE Trans. Inf. Theory* **2003**, *49*, 371–381.
48. Katti, S.; Rahul, H.; Hu, W.; Katabi, D.; Médard, M.; Crowcroft, J. XORs in the air: Practical wireless network coding. *IEEE/ACM Trans. Netw.* **2008**, *16*, 497–510.
49. Landau, L.; Lifshitz, E. *Statistical Physics*, 2nd ed.; Pergamon Press: Oxford, UK, 1970; Volume 5.
50. Wainwright, M.J.; Jordan, M.I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **2008**, *1*, 1–305.
51. Cipra, B.A. An introduction to the Ising model. *Am. Math. Mon.* **1987**, *94*, 937–959.
52. Sayed, A.H. *Adaptive Filters*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
53. El Gamal, A.; Kim, Y.H. *Network Information Theory*; Cambridge University Press: Cambridge, UK, 2011.