entropy

MDPI

Article

Comparing Information-Theoretic Measures of Complexity in Boltzmann Machines

Maxinder S. Kanwal [1,*], Joshua A. Grochow [2,3] and Nihat Ay [3,4,5]

1 Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA
2 Departments of Computer Science and Mathematics, University of Colorado, Boulder, CO 80309, USA; joshua.grochow@cs.colorado.edu
3 Santa Fe Institute, Santa Fe, NM 87501, USA; nay@mis.mpg.de
4 Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany
5 Faculty of Mathematics and Computer Science, University of Leipzig, 04009 Leipzig, Germany
* Correspondence: mkanwal@berkeley.edu; Tel.: +1-571-395-5242

Received: 30 April 2017; Accepted: 23 June 2017; Published: 3 July 2017

Abstract: In the past three decades, many theoretical measures of complexity have been proposed to help understand complex systems. In this work, for the first time, we place these measures on a level playing field, to explore the qualitative similarities and differences between them, and their shortcomings. Specifically, using the Boltzmann machine architecture (a fully connected recurrent neural network) with uniformly distributed weights as our model of study, we numerically measure how complexity changes as a function of network dynamics and network parameters. We apply an extension of one such information-theoretic measure of complexity to understand incremental Hebbian learning in Hopfield networks, a fully recurrent architecture model of autoassociative memory. In the course of Hebbian learning, the total information flow reflects a natural upward trend in complexity as the network attempts to learn more and more patterns.

Keywords: complexity; information integration; information geometry; Boltzmann machine; Hopfield network; Hebbian learning

1. Introduction

Many systems, across a wide array of disciplines, have been labeled "complex". The striking analogies between these systems [1,2] beg the question: What collective properties do complex systems share and what quantitative techniques can we use to analyze these systems as a whole? With new measurement techniques and ever-increasing amounts of data becoming available about larger and larger systems, we are in a better position than ever before to understand the underlying dynamics and properties of these systems.

While few researchers agree on a specific definition of a complex system, common terms used to describe complex systems include "emergence" and "self-organization", which characterize high-level properties in a system composed of many simpler sub-units. Often these sub-units follow local rules that can be described with much better accuracy than those governing the global system. Most definitions of complex systems include, in one way or another, the hallmark feature that the whole is more than the sum of its parts.

In the unified study of complex systems, a vast number of measures have been introduced to concretely quantify an intuitive notion of complexity (see, e.g., [3,4]). As Shalizi points out [4], among the plethora of complexity measures proposed, roughly, there are two main threads: those that build on the notion of Kolmogorov complexity and those that use the tools of Shannon's information theory. There are many systems for which the nature of their complexity seems

to stem either from logical/computational/descriptive forms of complexity (hence, Kolmogorov complexity) and/or from information-theoretic forms of complexity. In this paper, we focus on information-theoretic measures.

While the unified study of complex systems is the ultimate goal, due to the broad nature of the field, there are still many sub-fields within complexity science [1,2,5]. One such sub-field is the study of networks, and in particular, stochastic networks (broadly defined). Complexity in a stochastic network is often considered to be directly proportional to the level of stochastic interaction of the units that compose the network—this is where tools from information theory come in handy.

1.1. Information-Theoretic Measures of Complexity

Within the framework of considering stochastic interaction as a proxy for complexity, a few candidate measures of complexity have been developed and refined over the past decade. There is no consensus best measure, as each individual measure frequently captures some aspects of stochastic interaction better than others.

In this paper, we empirically examine four measures (described in detail later): (1) multi-information, (2) synergistic information, (3) total information flow, (4) geometric integrated information. Additional notable information-theoretic measures that we do not examine include those of Tononi et al., first proposed in [6] and most recently refined in [7], as a measure of consciousness, as well as similar measures of integrated information described by Barrett and Seth [8], and Oizumi et al. [9].

The term "humpology", first coined by Crutchfield [5], attempts to qualitatively describe a long and generally understood feature that a natural measure of complexity ought to have. In particular, as stochasticity varies from 0% to 100%, the structural complexity should be unimodal, with a maximum somewhere in between the extremes [10]. For a physical analogy, consider the spectrum of molecular randomness spanning from a rigid crystal (complete order) to a random gas (complete disorder). At both extremes, we intuitively expect no complexity: a crystal has no fluctuations, while a totally random gas has complete unpredictability across time. Somewhere in between, structural complexity will be maximized (assuming it is always finite).

We now describe the four complexity measures of interest in this study. We assume a compositional structure of the system and consider a finite set V of nodes. With each node $v \in V$, we associate a finite set $\mathbb{X}_v$ of states. In the prime example of this article, the Boltzmann machine, we have $V = \{1, \ldots, N\}$, and $\mathbb{X}_v = \{\pm 1\}$ for all v. For any subset $A \subseteq V$, we define the state set of all nodes in A as the Cartesian product $\mathbb{X}_A := \prod_{v \in A} \mathbb{X}_v$ and use the abbreviation $\mathbb{X} := \mathbb{X}_V$. In what follows, we want to consider stochastic processes in $\mathbb{X}$ and assign various complexity measures to these processes. With a probability vector $p(x)$, $x \in \mathbb{X}$, and a stochastic matrix $P(x, x')$, $x, x' \in \mathbb{X}$, we associate a pair (X, X') of random variables satisfying

$$p(x, x') := \Pr(X = x, X' = x') = p(x)P(x, x'), \qquad x, x' \in \mathbb{X}. \tag{1}$$

Obviously, any such pair of random variables satisfies $\Pr(X = x) = p(x)$, and $\Pr(X' = x' | X = x) = P(x, x')$ whenever $p(x) > 0$. As we want to assign complexity measures to transitions of the system state in time, we also use the more suggestive notation $X \to X'$ instead of (X, X'). If we iterate the transition, we obtain a Markov chain $X_n = (X_{n,v})_{v \in V}$, $n = 1, 2, \ldots$, in $\mathbb{X}$, with

$$p(x_1, x_2, \ldots, x_n) := \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p(x_1) \prod_{k=2}^{n} P(x_{k-1}, x_k), \qquad n = 1, 2, \ldots, \tag{2}$$

where, by the usual convention, the product on the right-hand side of this equation equals one if the index set is empty, that is, for $n = 1$. Obviously, we have $\Pr(X_1 = x) = p(x)$, and $\Pr(X_{n+1} = x' | X_n = x) = P(x, x')$ whenever $p(x) > 0$. Throughout the whole paper, we will assume that the probability vector p is stationary with respect to the stochastic matrix P. More precisely, we assume that for all $x' \in \mathbb{X}$ the following equality holds:

$$p(x') = \sum_{x \in \mathbb{X}} p(x) P(x, x').$$

With this assumption, we have $\Pr(X_n = x) = p(x)$, and the distribution of (X_n, X_{n+1}) does not depend on n. This will allow us to restrict attention to only one transition $X \to X'$. In what follows, we define various information-theoretic measures associated with such a transition.

1.1.1. Multi-Information, MI

The multi-information is a measure proposed by McGill [11] that captures the extent to which the whole is greater than the sum of its parts when averaging over time. For the above random variable X, it is defined as

$$MI(X) \triangleq \sum_{v \in V} H(X_v) - H(X), \tag{3}$$

where the Shannon entropy $H(X) = -\sum_{x \in \mathbb{X}} p(x) \log p(x)$. (Here, and throughout this article, we take logarithms with respect to base 2.) It holds that $MI(X) = 0$ if and only if all of the parts, X_i, are mutually independent.

1.1.2. Synergistic Information, SI

The synergistic information, proposed by Edlund et al. [12], measures the extent to which the (one-step) predictive information of the whole is greater than that of the parts. (For details related to the predictive information, see [13–15].) It builds on the multi-information by including the dynamics through time in the measure:

$$SI(X \to X') \triangleq I(X; X') - \sum_{v \in V} I(X_v; X'_v), \tag{4}$$

where $I(X; X')$ denotes the mutual information between X and X'. One potential issue with the synergistic information is that it may be negative. This is not ideal, as it is difficult to interpret a negative value of complexity. Furthermore, a preferred baseline minimum value of 0 serves as a reference point against which one can objectively compare systems.

The subsequent two measures (total information flow and geometric integrated information) have geometric formulations that make use of tools from information geometry. In information geometry, the Kullback–Leibler divergence (KL divergence) is used to measure the dissimilarity between two discrete probability distributions. Applied to our context, we measure the dissimilarity between two stochastic matrices P and Q with respect to p by

$$D^p_{KL}(P \| Q) = \sum_{x \in \mathbb{X}} p(x) \sum_{x' \in \mathbb{X}} P(x, x') \log \frac{P(x, x')}{Q(x, x')}. \tag{5}$$

For simplicity, let us assume that P and Q are strictly positive and that p is the stationary distribution of P. In that case, we do not explicitly refer to the stationary distribution p and simply write $D_{KL}(P \| Q)$. The KL divergence between P and Q can be interpreted by considering their corresponding Markov chains with distributions (2) (e.g., see [16] for additional details on this formulation). Denoting the chain of P by X_n, $n = 1, 2, \ldots$, and the chain of Q by Y_n, $n = 1, 2, \ldots$, with some initial distributions p_1 and q_1, respectively, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{x_1,x_2,\ldots,x_n}\Pr(X_1=x_1,X_2=x_2,\ldots,X_n=x_n)\log\frac{\Pr(X_1=x_1,X_2=x_2,\ldots,X_n=x_n)}{\Pr(Y_1=x_1,Y_2=x_2,\ldots,Y_n=x_n)}\\
&= \frac{1}{n}\left(\sum_{x_1}\Pr(X_1=x_1)\log\frac{\Pr(X_1=x_1)}{\Pr(Y_1=x_1)}+\right.\\
&\qquad\left.\sum_{k=1}^{n-1}\sum_{x}\Pr(X_k=x)\sum_{x'}\Pr(X_{k+1}=x'|X_k=x)\log\frac{\Pr(X_{k+1}=x'|X_k=x)}{\Pr(Y_{k+1}=x'|Y_k=x)}\right)\\
&= \frac{1}{n}\sum_{x}p_1(x)\log\frac{p_1(x)}{q_1(x)}+\frac{n-1}{n}\sum_{x}p(x)\sum_{x'}P(x,x')\log\frac{P(x,x')}{Q(x,x')}\\
&\overset{n\to\infty}{\to} D_{KL}(P\|Q).
\end{aligned}$$

We can use the KL divergence (5) to answer our original question—*To what extent is the whole greater than the sum of its parts?*—by comparing a system of interest to its most similar (least dissimilar) system whose whole is exactly *equal* to the sum of its parts. When comparing a transition P to Q using the KL divergence, one measures the amount of information lost when Q is used to approximate P. Hence, by constraining Q to be equal to the sum of its parts, we can then arrive at a natural measure of complexity by taking the *minimum* extent to which our distribution P is greater (in the sense that it contains more information) than some distribution Q, since Q represents a system of zero complexity. Formally, one defines a manifold $\mathcal{S}$, of so-called "split" systems, consisting of all those distributions that are equal to the sum of their parts, and then measures the minimum distance to that manifold:

$$Complexity(P) \triangleq \min_{Q\in\mathcal{S}} D_{KL}(P\|Q). \tag{6}$$

It is important to note here that there are many different viable choices of split manifold $\mathcal{S}$. This approach was first introduced by Ay for a general class of manifolds $\mathcal{S}$ [17]. Amari [18] and Oizumi et al. [19] proposed variants of this quantity as measures of information integration. In what follows, we consider measures of the form (6) for two different choices of $\mathcal{S}$.

1.1.3. Total Information Flow, IF

The total information flow, also known as the stochastic interaction, expands on the multi-information (like SI) to include temporal dynamics. Proposed by Ay in [17,20], the measure can be expressed by constraining Q to the manifold of distributions, $\mathcal{S}^{(1)}$, where there exists functions $f_v(x_v, x'_v)$, $v \in V$, such that Q is of the form:

$$Q(x,x') = Q((x_v)_{v\in V},(x'_v)_{v\in V}) = \frac{e^{\sum_{v\in V} f_v(x_v,x'_v)}}{Z(x)}, \tag{7}$$

where $Z(x)$ denotes the partition function that properly normalizes the distribution. Note that any stochastic matrix of this kind satisfies the property that $Q(x,x') = \prod_{v\in V}\Pr(X'_v = x'_v \mid X_v = x_v)$. This results in

$$\begin{aligned}
IF(X\to X') &\triangleq \min_{Q\in\mathcal{S}^{(1)}} D_{KL}(P\|Q) &(8)\\
&= \sum_{v\in V} H(X'_v \mid X_v) - H(X' \mid X). &(9)
\end{aligned}$$

The total information flow is non-negative, as are all measures that can be expressed as a KL divergence. One issue of note, as pointed out in [18,19], is that $IF(X\to X')$ can exceed $I(X;X')$. One can formulate the mutual information $I(X;X')$ as

$$I(X;X') = \min_{Q\in\mathcal{S}^{(2)}} D_{KL}(P\|Q), \tag{10}$$

where $\mathcal{S}^{(2)}$ consists of stochastic matrices Q that satisfy

$$Q(x, x') = Q((x_v)_{v \in V}, (x'_v)_{v \in V}) = \frac{e^{f_V(x')}}{Z(x)}, \tag{11}$$

for some function $f_V(x')$. Under this constraint, $Q(x, x') = \Pr(X' = x')$. In other words, all spatio-temporal interactions $X \to X'$ are lost. Thus, it has been postulated that no measure of information integration, such as the total information flow, should exceed the mutual information [9]. The cause of this violation in the total information flow is due to the fact that $IF(X \to X')$ quantifies same-time interactions in X' (due to the lack of an undirected edge in the output in Figure 1B). Consider, for instance, a stochastic matrix P that satisfies (11), $P(x, x') = p(x')$ for some probability vector p. In that case, we have $I(X; X') = 0$. However, (9) then reduces to the multi-information (3) of $X' = (X'_v)_{v \in V}$, which is a measure of stochastic dependence.

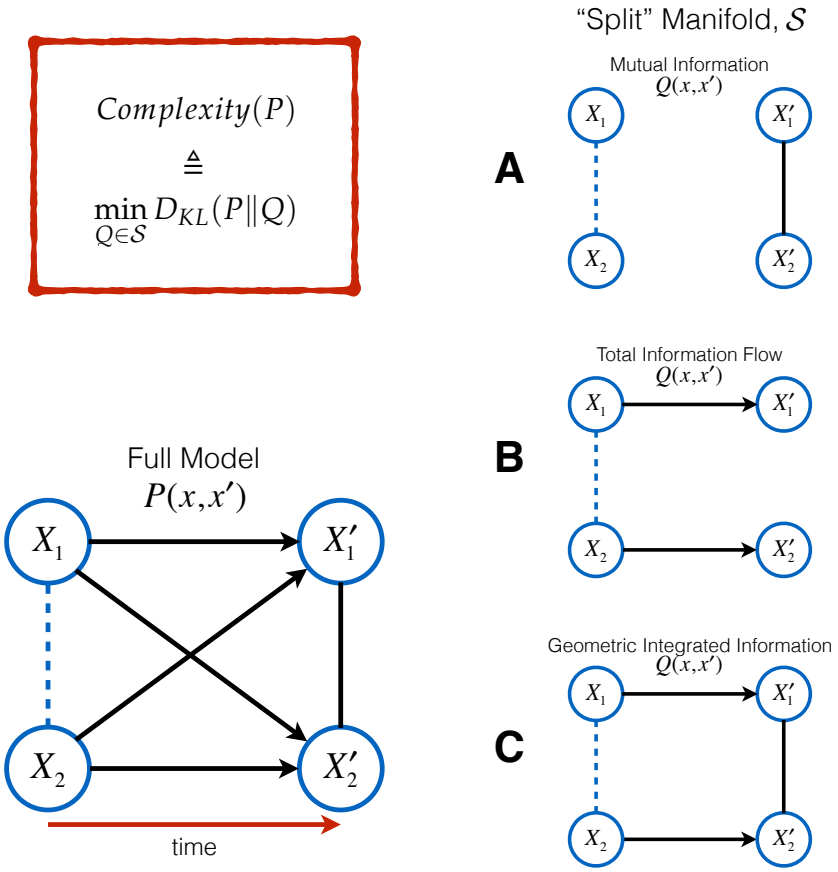


Figure 1. Using graphical models, we can visualize different ways to define the "split" constraint on manifold $\mathcal{S}$ in (6). Here, we consider a two-node network $X = (X_1, X_2)$ and its spatio-temporal stochastic interactions. (**A**) $I(X; X')$ uses constraint (11). (**B**) $IF(X \to X')$ uses constraint (7). (**C**) $\Phi_G(X \to X')$ uses constraint (13). Dashed lines represent correlations that either may or may not be present in the input distribution p. We do not represent these correlations with solid lines in order to highlight (with solid lines) the structure imposed on the stochastic matrices. Adapted and modified from [19].

1.1.4. Geometric Integrated Information, Φ_G

In order to obtain a measure of information integration that does not exceed the mutual information $I(X; X')$, Amari [18] (Section 6.9) defines $\Phi_G(X \to X')$ as

$$\Phi_G(X \to X') \triangleq \min_{Q \in \mathcal{S}^{(3)}} D_{KL}(P \| Q), \tag{12}$$

where $\mathcal{S}^{(3)}$ contains not only the split matrices (7), but also those matrices that satisfy (11). More precisely, the set $\mathcal{S}^{(3)}$ consists of all stochastic matrices for which there exists functions $f_v(x_v, x'_v)$, $v \in V$, and $f_V(x')$ such that

$$Q(x, x') = Q((x_v)_{v \in V}, (x'_v)_{v \in V}) = \frac{e^{\sum_{v \in V} f_v(x_v, x'_v) + f_V(x')}}{Z(x)}. \tag{13}$$

Here, Q belongs to the set of matrices where only time-lagged interactions are removed. Note that the manifold $\mathcal{S}^{(3)}$ contains $\mathcal{S}^{(1)}$, the model of split matrices used for IF, as well as $\mathcal{S}^{(2)}$, the manifold used for the mutual information. This measure thus satisfies both postulates that SI and IF only partially satisfy:

$$0 \leq \Phi_G(X \to X') \leq I(X; X'). \tag{14}$$

However, unlike $IF(X \to X')$, there is no closed-form expression to use when computing $\Phi_G(X \to X')$. In this paper, we use the iterative scaling algorithm described in [21] (Section 5.1) to compute $\Phi_G(X \to X')$ for the first time in concrete systems of interest.

Note that, in defining $\Phi_G(X \to X')$, the notion of a split model used by Amari [18] is related, but not identical, to that used by Oizumi et al. [19]. The manifold considered in the latter work is defined in terms of conditional independence statements and forms a curved exponential family.

In the remainder of this article, we also use the shorthand notation MI, SI, IF, and Φ_G, without explicit reference to X and X', as already indicated in each measure's respective subsection heading. We also use I as shorthand for the mutual information.

1.2. Boltzmann Machine

In this paper, we look at the aforementioned candidate measures in a concrete system in order to gain an intuitive sense of what is frequently discussed at a heavily theoretical and abstract level. Our system of interest is the Boltzmann machine (a fully-recurrent neural network with sigmoidal activation units).

We parameterize a network of N binary nodes by $W \in \mathbb{R}^{N \times N}$, which denotes the connectivity matrix of weights between each directed pair of nodes. Each node i takes a value $X_i \in \{\pm 1\}$, and updates to $X'_i \in \{\pm 1\}$ according to:

$$\Pr(X'_i = +1 \mid X) = sigmoid\Big(-2\beta \sum_{j=1}^{N} w_{ji} \cdot X_j \Big), \tag{15}$$

where $sigmoid(t) = \frac{1}{1+e^{-t}}$, β denotes a global inverse-temperature parameter, and w_{ji} denotes the directed weight from X_j to X_i.

This stochastic update rule implies that every node updates probabilistically according to a weighted sum of the node's parents (or inputs), which, in the case of our fully recurrent neural network, is every node in the network. Every node i has some weight, w_{ij}, with which it influences node j on the next update. As the weighted sum of the inputs to a node becomes more positive, the likelihood of that node updating to the state $+1$ increases. The opposite holds true as the weighted sum becomes more negative, as seen in Figure 2. The weights between nodes are a set of parameters that we are free to tune in the network.

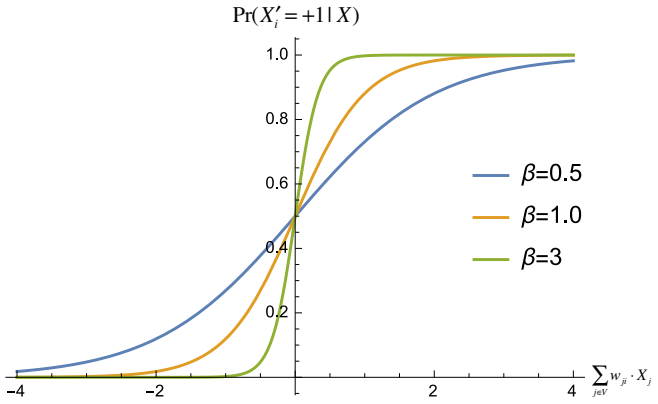


Figure 2. The sigmoidal update rule as a function of the inverse-global temperature: As β increases, the stochastic update rule becomes closer to the deterministic one given by a step function.

The second tunable parameter in our network is β, commonly known as the global inverse-temperature of the network. β effectively controls the extent to which the system is influenced by random noise: it quantifies the system's deviation from deterministic updating. In networks, the noise level directly correlates with what we call the "pseudo-temperature" T of the network, where $T = \frac{1}{\beta}$. To contextualize what T might represent in a real-life complex system, consider the example of a biological neural network, where we can think of the pseudo-temperature as a parameter that encompasses all of the variables (beyond just a neuron's synaptic inputs) that influence whether a neuron fires or not in a given moment (e.g., delays in integrating inputs, random fluctuations from the release of neurotransmitters in vesicles, firing of variable strength). As $\beta \to 0$ ($T \to \infty$), the interactions are governed entirely by randomness. On the other hand, as $\beta \to \infty$ ($T \to 0$), the nodal inputs takeover as the only factor in determining the subsequent states of the units—the network becomes deterministic rather than stochastic.

This sigmoidal update rule is commonly used as the nonlinearity in the nodal activation function in stochastic neural networks for reasons coming from statistical mechanics: It arises as a direct consequence of the Boltzmann–Gibbs distribution when assuming pairwise interactions (similar to Glauber dynamics on the Ising model), as explained in, for example, [22] (Chapter 2 and Appendix A). As a consequence of this update rule, for finite β, there is always a unique stationary distribution on the stochastic network state space.

2. Results

What follows are plots comparing and contrasting the four introduced complexity measures in their specified settings. The qualitative trends shown in the plots empirically hold regardless of network size; a 5-node network was used to generate the plots below.

In Figure 3a, we see that when weights are uniformly distributed between 0 and 1, IF and Φ_G are very similar qualitatively, with the additional property that $\Phi_G \leq IF$, which directly follows from $\mathcal{S}^{(1)} \subseteq \mathcal{S}^{(3)}$. MI monotonically increases, which contradicts the intuition prescribed by humpology. Finally, SI is peculiar in that it is not lower-bounded by 0. This makes for difficult interpretation: what does a negative complexity mean as opposed to zero complexity? Furthermore, in Figure 3b, we see that Φ_G satisfies constraint (14), with the mutual information in fact upper bounding both IF and Φ_G.

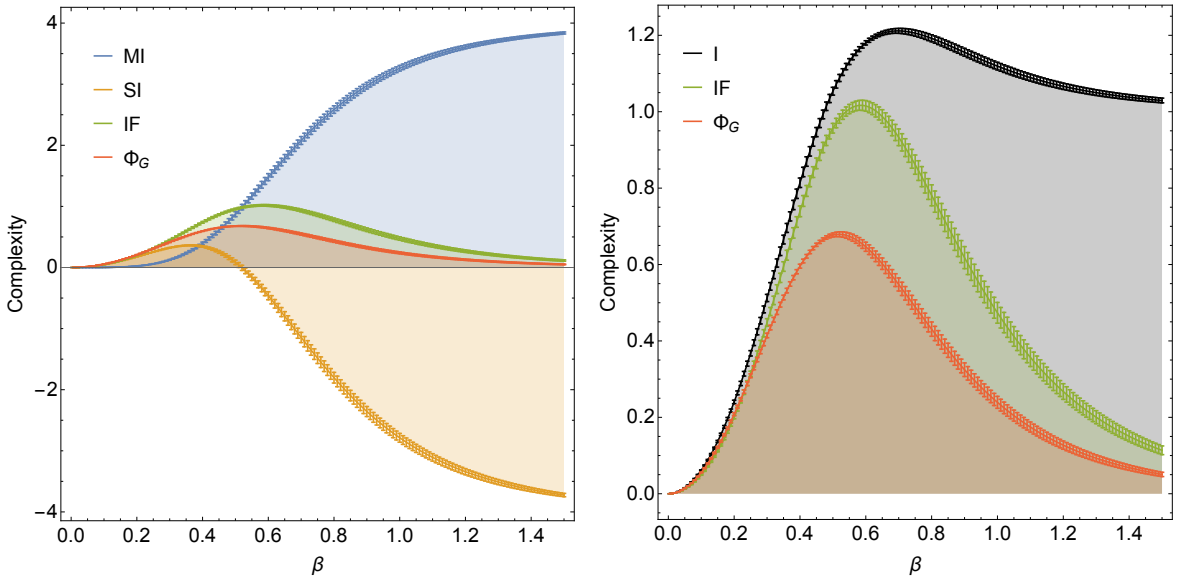


Figure 3. (**a**) measures of complexity when using random weight initializations sampled uniformly between 0 and 1 (averaged over 100 trials, with error bars); (**b**) the mutual information I upper bounds IF and Φ_G when using random weight initializations sampled uniformly between 0 and 1 (averaged over 100 trials, with error bars).

It is straightforward to see the symmetry between selecting weights uniformly between 0 and $+1$ and between -1 and 0, hence the above results represent both scenarios.

When we allow for both positive and negative weights, however, about as frequently as we observe the above behavior, we observe qualitatively different behavior as represented in Figure 4. Physically, these results correspond to allowing for mutual excitation and inhibition in the same network.

In Figure 4a, surprisingly, we see that in one instance of mixed weights, IF monotonically increases (like MI in Figure 3a), a departure from humpology intuition. Meanwhile, Φ_G behaves qualitatively differently, such that $\Phi_G \to 0$ as $\beta \to \infty$. In Figure 4b, we see an instance where all measures limit to some non-zero value as $\beta \to \infty$. Finally, in Figure 4c, we see an instance where IF exceeds I while Φ_G satisfies constraint (14), despite the common unimodality of both measures.

An overly simplistic interpretation of the idea that humpology attempts to capture may lead one to believe that Figure 4b is a negative result discrediting all four measures. We claim, however, that this result suggests that the simple humpology intuition described in Section 1.1 needs additional nuance when applied to quantifying the complexity of dynamical systems. In Figure 4b, we observe a certain richness to the network dynamics, despite its deterministic nature. A network dynamics that deterministically oscillates around a non-trivial attractor is not analogous to the "frozen" state of a rigid crystal (no complexity). Rather, one may instead associate the crystal state with a network whose dynamics is the identity map, which can indeed be represented by a split stochastic matrix. Therefore, whenever the stochastic matrix P converges to the identity matrix (the "frozen" matrix) for $\beta \to \infty$, the complexity will asymptote to zero (as in Figure 3b). In other words, for dynamical systems, a "frozen" system is exactly that: a network dynamics that has settled into a single *fixed-point* dynamics. Consequently, in our results, as $\beta \to \infty$, we should expect that the change in complexity depends on the dynamics that the network is settling into as it becomes deterministic, and the corresponding richness (e.g., number of attractors and their lengths) of that asymptotic dynamics.

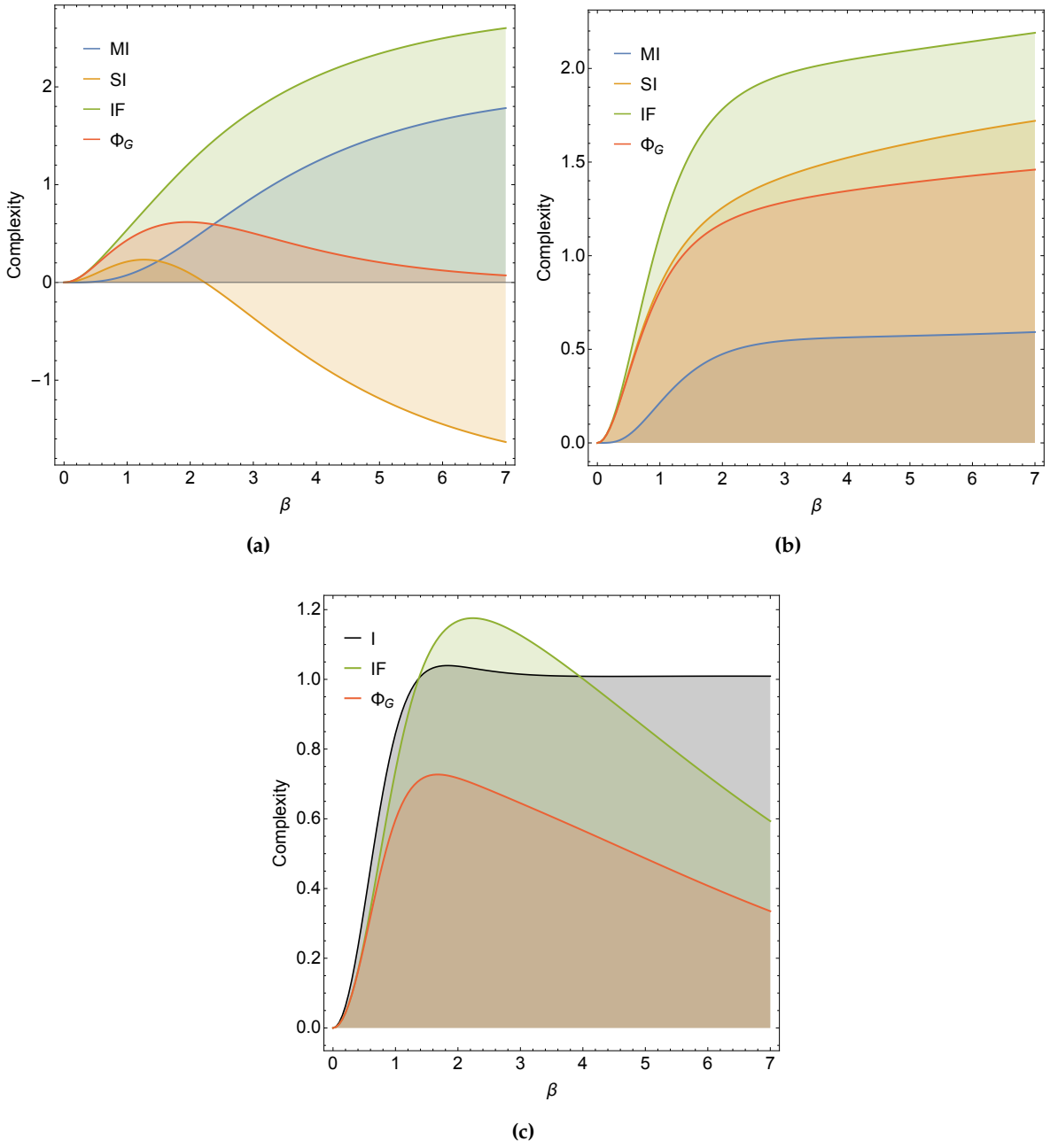


Figure 4. Measures of complexity in single instances of using random weight initializations sampled uniformly between -1 and 1. (**a**) scenario 1; (**b**) scenario 2; (**c**) constraint (14).

So far, it may seem to be the case that Φ_G is without flaw; however, there are shortcomings that warrant further study. In particular, in formulating Φ_G, the undirected output edge in Figure 5B (purple) was deemed necessary to avoid quantifying external influences to the system that IF would consider as intrinsic information flow. Yet, in the model studied here—the Boltzmann machine—there are no such external influences (i.e., $Y = 0$ in Figure 5), so this modification should have no effect on distinguishing between Φ_G and IF in our setting. More precisely, a full model that lacks an undirected output edge at the start should not lead to a "split"-projection that incorporates such an edge. However, this is not generally true for the projection that Φ_G computes because the undirected output edge present in the split model will in fact capture causal interactions *within* the system by deviously interpreting them as same-time interactions in the output (Figure 5). This counterintuitive phenomenon suggests that we should have preferred IF to be precisely equal to its ideal form Φ_{ideal}

in the case of the Boltzmann machine, and yet, almost paradoxically, this would imply that the improved form would still violate constraint (14). This puzzling conundrum begs further study of how to properly disentangle external influences when attempting to strictly quantify the intrinsic causal interactions.

The preceding phenomenon, in fact, also calls into question the very postulate that the mutual information ought to be an upper bound on information integration. As we see in Figure 5A, the undirected output edge used in the "split"-projection for computing the mutual information I is capable of producing the very same problematic phenomenon. Thus, the mutual information does not fully quantify the total causal influences intrinsic to a system. In fact, the assumption itself that I quantified the total intrinsic causal influences was based on the assumption that one can distinguish between intrinsic and extrinsic influences in the first place, which may not be the case.

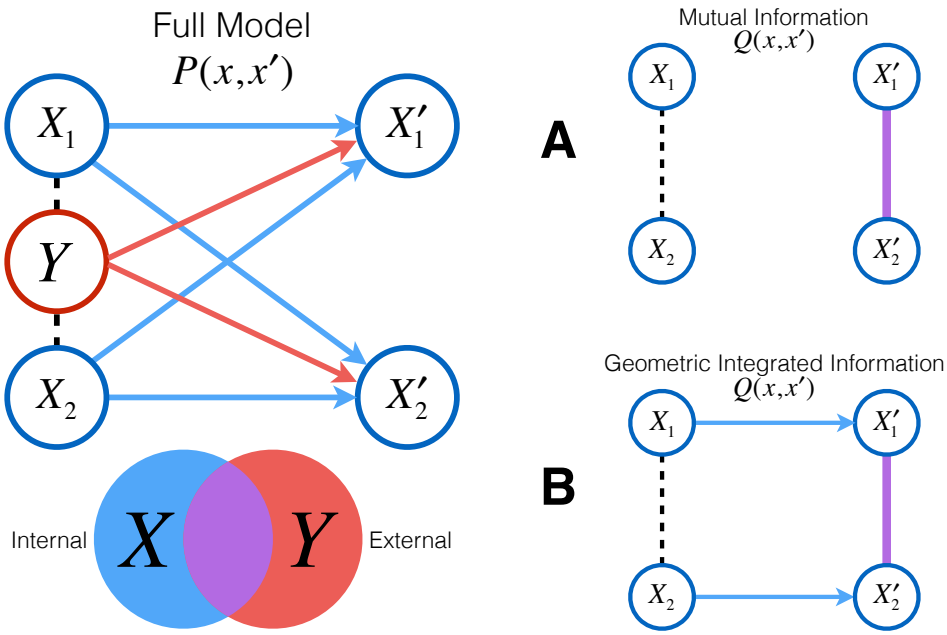


Figure 5. A full model (left) can have both intrinsic (blue) and extrinsic (red) causal interactions contributing to its overall dynamics. Split models (**A**,**B**) formulated with an undirected output edge (purple) attempt to exclusively quantify extrinsic causal interactions (so as to strictly preserve intrinsic causal interactions after the "split"-projection). However, the output edge can end up explaining away interactions from *both* external factors Y *and* (some) internal factors X (red + blue = purple). As a result, such a family of split models does not properly capture the *total* intrinsic causal interactions present in a system.

3. Application

In this section, we apply one of the preceding measures (IF) and examine its dynamics during network learning. We wish to exemplify the insights that one can gain by exploring measures of complexity in a more general sense. The results presented in Section 2 showed the promising nature of information-geometric formulations of complexity, such as IF and Φ_G. Here, however, we restrict ourselves to studying IF as a first step due to the provable properties of its closed-form expression that we are able to exploit to study it in greater depth in the context of autoassociative memory networks. It would be useful to extend this analysis to Φ_G, but this is beyond the scope of this work.

Autoassociative memory in a network is a form of "collective computation" where, given an incomplete input pattern, the network can accurately recall a previously stored pattern by evolving from the input to the stored pattern. For example, a pattern might be a binary image, in which each pixel in the image corresponds to a node in the network with a value in $\{-1, +1\}$. In this

case, an autoassociative memory model with a stored image could then take as input a noisy version of the stored image and accurately recall the fully denoised original image. This differs from a "serial computation" approach to the same problem where one would simply store the patterns in a database and, when given an input, search all images in the database for the most similar stored image to output.

One mechanism by which a network can achieve collective computation has deep connections to concepts from statistical mechanics (e.g., the Ising model, Glauber dynamics, Gibbs sampling). This theory is explained in detail in [22]. The clever idea behind autoassociative memory models heavily leverages the existence of an energy function (sometimes called a Lyapunov function) to govern the evolution of the network towards a locally minimal energy state. Thus, by engineering the network's weighted edges such that local minima in the energy function correspond to stored patterns, one can show that if an input state is close enough (in Hamming distance) to a desired stored state, then the network will evolve towards the correct lower-energy state, which will in fact be a stable fixed point of the network.

The above, however, is only true up to a limit. A network can only store so many patterns before it becomes saturated. As more and more patterns are stored, various problems arise such as desirable fixed points becoming unstable optima, as well as the emergence of unwanted fixed points in the network that do not correspond to any stored patterns (i.e., spin glass states).

In 1982, Hopfield put many of these ideas together to formalize what is today known as the Hopfield model, a fully recurrent neural network capable of autoassociative memory. Hopfield's biggest contribution in his seminal paper was assigning an energy function to the network model:

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} X_i X_j. \tag{16}$$

For our study, we assume that we are storing random patterns in the network. In this scenario, Hebb's rule (Equation (17)) is a natural choice for assigning weights to each connection between nodes in the network such that the random patterns are close to stable local minimizers of the energy function.

Let $\{\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(T)}\}$ denote the set of N-bit binary patterns that we desire to store. Then, under Hebb's rule, the weight between nodes i and j should be assigned as follows:

$$w_{ij} = \frac{1}{T}\sum_{\mu=1}^{T} \xi_i^{(\mu)} \xi_j^{(\mu)}, \tag{17}$$

where $\xi_i^{(\mu)}$ denotes the ith-bit of pattern $\xi^{(\mu)}$. Notice that all weights are symmetric, $w_{ij} = w_{ji}$.

Hebb's rule is frequently used to model learning, as it is both *local* and *incremental*—two desirable properties of a biologically plausible learning rule. Hebb's rule is local because weights are set based strictly on local information (i.e., the two nodes that the weight connects) and is incremental because new patterns can be learned one at a time without having to reconsider information from already learned patterns. Hence, under Hebb's rule, training a Hopfield network is relatively simple and straightforward.

The update rule that governs the network's dynamics is the same sigmoidal function used in the Boltzmann machine described in Section 1.2. We will have this update rule take effect synchronously for all nodes (Note: Hopfield's original model was described in the asynchronous, deterministic case but can also be studied more generally.):

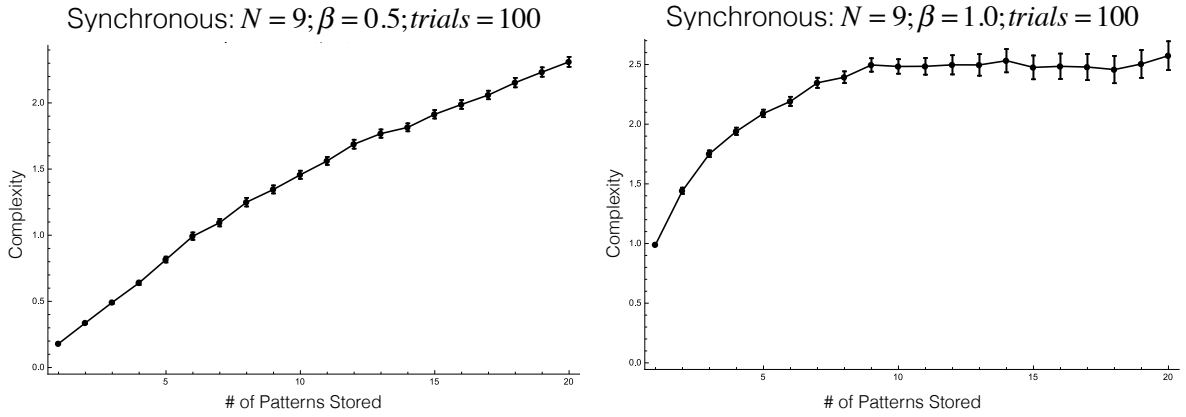
$$\Pr(X_i' = +1 \mid X) = \frac{1}{1 + e^{-2\beta \sum_{j \in V} X_j \cdot w_{ji}}}. \tag{18}$$

At finite β, our Hopfield model obeys a stochastic sigmoidal update rule. Thus, there exists a unique and strictly positive stationary distribution of the network dynamics.

Here, we study *incremental* Hebbian learning, in which multiple patterns are stored in a Hopfield network in succession. We use total information flow (Section 1.1.3) to explore how incremental Hebbian learning changes complexity, or more specifically, how the complexity relates to the number of patterns stored.

Before continuing, we wish to make clear upfront an important disclaimer: the results that we describe are qualitatively different when one uses asynchronous dynamics instead of synchronous, as we use here. With asynchronous dynamics, no significant overall trend manifests, but other phenomena emerge in need of further exploration.

When we synchronously update nodes, we see very interesting behavior during learning: incremental Hebbian learning appears to increase complexity, on average (Figure 6a,b). The dependence on β is not entirely clear, but as one can infer from Figure 6a,b, it appears that increasing β increases the magnitude of the average complexity while learning, while also increasing the variance of the complexity. Thus, as β increases, the average case becomes more and more unrepresentative of the individual cases of incremental Hebbian learning.



(a) $\beta = \frac{1}{2}$ **(b)** $\beta = 1$

Figure 6. Incremental Hebbian learning in a 9-node stochastic Hopfield network with synchronous updating (averaged over 100 trials of storing random 9-bit patterns). (**a**) $\beta = \frac{1}{2}$; (**b**) $\beta = 1$.

We can also study the deterministic version of the Hopfield model. This corresponds to letting $\beta \to \infty$ in the stochastic model. With a deterministic network, many stationary distributions on the network dynamics may exist, unlike in the stochastic case. As discussed above, if we want to recall a stored image, we would like for that image to be a fixed point in the network (corresponding to a stationary distribution equal to the Dirac measure at that state). Storing multiple images corresponds to the desire to have multiple Dirac measures acting as stationary distributions of the network. Furthermore, in the deterministic setting, the nodal update rule becomes a step rather than a sigmoid function.

Without a unique stationary distribution in the deterministic setting, we must decide how to select an input distribution to use in calculating the complexity. If there are multiple stationary distributions in a network, not all starting distributions on the network eventually lead to a single stationary distribution (as was the case in the stochastic model), but instead the stationary distribution that the network eventually reaches is sensitive to the initial state of the network. When there are multiple stationary distributions, there are actually infinitely many stationary distributions, as any convex combination of stationary distributions is also stationary. If there exist N orthogonal stationary distributions of a network, then there is in fact an entire $(N-1)$-simplex of stationary distributions, any of which could be used as the input distribution for calculating the complexity.

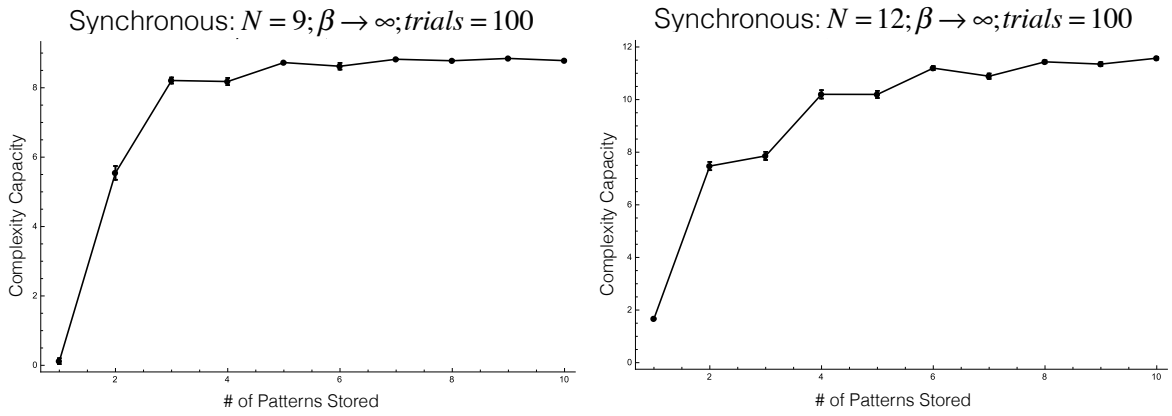
In order to address this issue, it is fruitful to realize that the complexity measure we are working with is concave with respect to the input distribution (Theorem A1 in Appendix A). As a function of the input distribution, there is thus an "apex" to the complexity. In other words, it is a unique local maximum of the complexity function, which is also therefore a global maximum (but not necessarily a unique maximizer since the complexity is not strictly concave). This means that the optimization problem of finding the supremum over the entire complexity landscape with respect to the input distribution is relatively simple and can be viably achieved via standard gradient-based methods.

We can naturally define a new quantity to measure complexity of a stochastic matrix P in this setting, the *complexity capacity*:

$$C_{cap}(X \to X' \mid P) \triangleq \max_{p} C(X \to X' \mid p, P), \tag{19}$$

where the maximum is taken over all stationary distributions p of P. Physically, the complexity capacity measures the *maximal* extent—over possible input distributions—to which the whole is more than the sum of its parts. By considering the entire convex hull of stationary input distributions and optimizing for complexity, we can find this unique maximal value and use it to represent the complexity of a network with multiple stationary distributions.

Again, in the synchronous-update setting, we see incremental Hebbian learning increases complexity *capacity* (Figure 7a,b). It is also worth noting that the complexity capacity in this setting is limiting towards the absolute upper bound on the complexity, which can never exceed the number of binary nodes in the network. Physically, this corresponds to each node attempting to store one full bit (the most information a binary node can store), and all of this information flowing through the network between time-steps, as more and more patterns are learned. This limiting behavior of the complexity capacity towards a maximum (as the network saturates with information) is more gradual as the size of the network increases. This observed behavior matches the intuition that larger networks should be able to store more information than smaller networks.



(a) $N = 9$ **(b)** $N = 12$

Figure 7. Incremental Hebbian learning in an N-node deterministic ($\beta \to \infty$) Hopfield network with synchronous updating (averaged over 100 trials of storing random N-bit patterns). (**a**) $N = 9$; (**b**) $N = 12$.

4. Conclusions

In summary, we have seen four different measures of complexity applied in concrete, parameterized systems. We observed that the synergistic information was difficult to interpret on its own due to the lack of an intuitive lower bound on the measure. Building off of the primitive multi-information, the total information flow and the geometric integrated information were closely related, frequently (but not always) showing the same qualitative behavior. The geometric integrated information satisfies the additional postulate (14) stating that a measure of complexity should not exceed the temporal

mutual information, a property that the total information flow frequently violated in the numerical experiments where connection weights were allowed to be both negative and positive. The geometric integrated information was recently proposed to build on and correct the original flaws in the total information flow, which it appears to have done quite singularly based on the examination in the present study. While the geometric integrated information is a step in the right direction, further study is needed to properly disentangle external from internal causal influences that contribute to network dynamics (see final paragraphs of Section 2). Nonetheless, it is encouraging to see a semblance of convergence with regards to quantifying complexity from an information-theoretic perspective.

Acknowledgments: The authors would like to thank the Santa Fe Institute (SFI) National Science Foundation (NSF) Research Experiences for Undergraduates program (NSF grant #1358567), where this work began, and the Max Planck Institute for Mathematics in the Sciences (German Research Foundation (DFG) Priority Program 1527 Autonomous Learning), where this work continued. J.A.G. was supported by an SFI Omidyar Fellowship during this work.

Author Contributions: Nihat Ay and Joshua A. Grochow proposed the research. Maxinder S. Kanwal carried out most of the research and took the main responsibility for writing the article. All authors contributed to joint discussions. All authors read and approved the final manuscript.

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

Appendix A

Theorem A1 (Concavity of $IF(X \to X')$)**.** *The complexity measure*

$$IF(X \to X') \triangleq \sum_{v \in V} H(X'_v \mid X_v) - H(X' \mid X),$$

is concave with respect to the input distribution $p(x) = \Pr(X = x)$, $x \in \mathbb{X}$, *for stochastic matrix P fixed.*

Note that in the definition of the *complexity capacity* (19), we take the supremum over all *stationary* input distributions. Since such distributions form a convex subset of the set of all input distributions, concavity of IF is preserved by the corresponding restriction.

Proof. The proof of the above statement follows from first rewriting the complexity measure in terms of a negative KL divergence between two distributions both affine with respect to the input distribution, and then using the fact that the KL divergence is convex with respect to a pair of distributions (see [23], Chapter 2) to demonstrate that the complexity measure is indeed concave.

Let P denote the fixed stochastic matrix governing the evolution of $X \to X'$.

Let p denote the input distribution on the states of X.

First, note that the domain of p forms a convex set: for an N-unit network, the set of all valid distributions p forms an $(N-1)$-simplex.

Next, we expand IF:

$$\begin{aligned} IF(X \to X') &= \sum_{v \in V} H(X'_v \mid X_v) - H(X' \mid X) \\ &= -\sum_{v \in V} \left(\sum_{x_v \in \mathbb{X}_v} \Pr(X_v = x_v) \sum_{x'_v \in \mathbb{X}_v} \Pr(X'_v = x'_v \mid X_v = x_v) \cdot \log \Pr(X'_v = x'_v \mid X_v = x_v) \right) \\ &\quad + \sum_{x \in \mathbb{X}} \Pr(X = x) \sum_{x' \in \mathbb{X}} \Pr(X' = x' \mid X = x) \cdot \log \Pr(X' = x' \mid X = x). \end{aligned}$$

Notice that the expanded expression for $H(X' \mid X)$ is affine in the input distribution $p(x)$, since the terms $\Pr(X' = x' \mid X = x)$ are just constants given by $P(x, x')$. Hence, $-H(X' \mid X)$ is concave, and all that is left to show is that the expansion of $H(X'_v \mid X_v)$ is also concave for all $v \in V$:

$$
\begin{aligned}
H(X'_v \mid X_v) &= -\sum_{x_v \in \mathbb{X}_v} \Pr(X_v = x_v) \sum_{x'_v \in \mathbb{X}_v} \Pr(X'_v = x'_v \mid X_v = x_v) \cdot \log \Pr(X'_v = x'_v \mid X_v = x_v) \\
&= -\sum_{x_v \in \mathbb{X}_v} \sum_{x'_v \in \mathbb{X}_v} \Pr(X'_v = x'_v, X_v = x_v) \cdot \log \frac{\Pr(X'_v = x'_v, X_v = x_v)}{\Pr(X_v = x_v)} \\
&= -\sum_{x_v \in \mathbb{X}_v} \sum_{x'_v \in \mathbb{X}_v} \Pr(X'_v = x'_v, X_v = x_v) \cdot \log \frac{\Pr(X'_v = x'_v, X_v = x_v)}{\Pr(X_v = x_v)} \\
&\quad + \log \frac{1}{|\mathbb{X}_v|} - \log \frac{1}{|\mathbb{X}_v|} \\
&= -\sum_{x_v \in \mathbb{X}_v} \sum_{x'_v \in \mathbb{X}_v} \Pr(X'_v = x'_v, X_v = x_v) \cdot \log \frac{\Pr(X'_v = x'_v, X_v = x_v)}{\Pr(X_v = x_v)} \\
&\quad + \sum_{x_v \in \mathbb{X}_v} \sum_{x'_v \in \mathbb{X}_v} \Pr(X'_v = x'_v, X_v = x_v) \log \frac{1}{|\mathbb{X}_v|} \\
&\quad - \log \frac{1}{|\mathbb{X}_v|} \\
&= -\sum_{x_v \in \mathbb{X}_v} \sum_{x'_v \in \mathbb{X}_v} \left(\Pr(X'_v = x'_v, X_v = x_v) \cdot \log \frac{\Pr(X'_v = x'_v, X_v = x_v)}{\frac{1}{|\mathbb{X}_v|} \cdot \Pr(X_v = x_v)} \right) - \log \frac{1}{|\mathbb{X}_v|}.
\end{aligned}
$$

Ignoring the constant $-\log \frac{1}{|\mathbb{X}_v|}$, as this does not change the concavity of the expression, we can rewrite the summation as

$$
= -\sum_{x_v \in \mathbb{X}_v} \sum_{x'_v \in \mathbb{X}_v} \left(\left(\sum_{x_r \in \mathbb{X}_{V \setminus v}} \Pr(X'_v = x'_v \mid X_v = x_v, X_{V \setminus v} = x_r) \cdot \Pr(X_v = x_v, X_{V \setminus v} = x_r) \right) \cdot \log \frac{\sum_{x_r \in \mathbb{X}_{V \setminus v}} \Pr(X'_v = x'_v | X_v = x_v, X_{V \setminus v} = x_r) \cdot \Pr(X_v = x_v, X_{V \setminus v} = x_r)}{\frac{1}{|\mathbb{X}_v|} \cdot \sum_{x_r \in \mathbb{X}_{V \setminus v}} \Pr(X_v = x_v, X_{V \setminus v} = x_r)} \right),
$$

where $X_{V \setminus v}$ denotes the state of all nodes excluding X_v. This expansion has made use of the fact that $\Pr(X'_v = x'_v, X_v = x_v) = \sum_{x_r \in \mathbb{X}_{V \setminus v}} \Pr(X'_v = x'_v \mid X_v = x_v, X_{V \setminus v} = x_r) \cdot \Pr(X_v = x_v, X_{V \setminus v} = x_r)$ and $\Pr(X_v = x_v) = \sum_{x_r \in \mathbb{X}_{V \setminus v}} \Pr(X_v = x_v, X_{V \setminus v} = x_r)$.

The constant $\Pr(X'_v = x'_v \mid X_v = x_v, X_{V \setminus v} = x_r) = \Pr(X'_v = x'_v \mid X = (x_v, x_r))$ can be computed directly as a marginal over the stochastic matrix P. Furthermore, the constant $\Pr(X_v = x_v, X_{V \setminus v} = x_r) = \Pr(X = (x_v, x_r))$ comes directly from the input distribution p, making the entire expression for $\Pr(X'_v = x'_v, X_v = x_v)$ affine with respect to the input distribution.

Finally, we get

$$
\begin{aligned}
&= -D_{KL} \left(\sum_{x_r \in \mathbb{X}_{V \setminus v}} \Pr(X'_v = x'_v \mid X_v = x_v, X_{V \setminus v} = x_r) \cdot \Pr(X_v = x_v, X_{V \setminus v} = x_r) \;\middle\|\; \frac{1}{|\mathbb{X}_v|} \cdot \sum_{x_r \in \mathbb{X}_{V \setminus v}} \Pr(X_v = x_v, X_{V \setminus v} = x_r) \right) \\
&= -D_{KL} \left(\Pr(X'_v = x'_v, X_v = x_v) \;\middle\|\; \frac{1}{|\mathbb{X}_v|} \cdot \Pr(X_v = x_v) \right),
\end{aligned}
$$

the KL divergence between two distributions, both of which have been written so as to explicitly show them as affine in the input distribution p, and then simplified to show that both are valid joint distributions over the states on the pair (X'_v, X_v). Thus, the overall expression is concave with respect to the input distribution. □

References

1. Miller, J.H.; Page, S.E. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*; Princeton University Press: Princeton, NJ, USA, 2007.
2. Mitchell, M. *Complexity: A Guided Tour*; Oxford University Press: Oxford, UK, 2009.
3. Lloyd, S. Measures of Complexity: A Non-Exhaustive List. Available online: http://web.mit.edu/esd.83/www/notebook/Complexity.PDF (accessed on 16 July 2016).
4. Shalizi, C. Complexity Measures. Available online: http://bactra.org/notebooks/complexity-measures.html (accessed on 16 July 2016).
5. Crutchfield, J.P. Complex Systems Theory? Available online: http://csc.ucdavis.edu/~chaos/chaos/talks/CSTheorySFIRetreat.pdf (accessed on 16 July 2016).
6. Tononi, G.; Sporns, O.; Edelman, G.M. A Measure for Brain Complexity: Relating Functional Segregation and Integration in the Nervous System. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 5033–5037.
7. Oizumi, M.; Albantakis, L.; Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput. Biol.* **2014**, *10*, 1–25.
8. Barrett, A.B.; Seth, A.K. Practical Measures of Integrated Information for Time-Series Data. *PLoS Comput. Biol.* **2011**, *7*, 1–18.
9. Oizumi, M.; Amari, S.I.; Yanagawa, T.; Fujii, N.; Tsuchiya, N. Measuring Integrated Information from the Decoding Perspective. *PLoS Comput. Biol.* **2016**, *12*, 1–18.
10. Gell-Mann, M. *The Quark and the Jaguar: Adventures in the Simple and the Complex*; St. Martin's Griffin: New York, NY, USA, 1994.
11. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
12. Edlund, J.A.; Chaumont, N.; Hintze, A.; Christof Koch, G.T.; Adami, C. Integrated Information Increases with Fitness in the Evolution of Animats. *PLoS Comput. Biol.* **2011**, *7*, 1–13.
13. Bialek, W.; Nemenman, I.; Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **2001**, *13*, 2409–2463.
14. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938.
15. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos* **2003**, *13*, 25–54.
16. Nagaoka, H. The exponential family of Markov chains and its information geometry. In Procedings of the 28th Symposium on Information Theory and Its Applications (SITA2005), Okinawa, Japan, 20–23 November 2005; pp. 601–604.
17. Ay, N. Information Geometry on Complexity and Stochastic Interaction. *MPI MIS PREPRINT 95* **2001**. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.6974 (accessed on 3 July 2017).
18. Amari, S. *Information Geometry and Its Applications*; Springer: Chiyoda, Japan, 2016.
19. Oizumi, M.; Tsuchiya, N.; Amari, S.I. Unified framework for information integration based on information geometry. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 14817–14822.
20. Ay, N. Information Geometry on Complexity and Stochastic Interaction. *Entropy* **2015**, *17*, 2432–2458.
21. Csiszár, I.; Shields, P. Information Theory and Statistics: A Tutorial. *Found. Trends® Commun. Inf. Theory* **2004**, *1*, 417–528.
22. Hertz, J.; Krogh, A.; Palmer, R.G. *Introduction to the Theory of Neural Computation*; Perseus Publishing: New York, NY, USA, 1991.
23. Cover, T.M.; Thomas, J.A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*; Wiley-Interscience: Hoboken, NJ, USA, 2006.