Entropy **2015**, *17*, 2432-2458; doi:10.3390/e17042432

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Information Geometry on Complexity and Stochastic Interaction

Nihat Ay [1,2,3]

[1] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany; E-Mail: nay@mis.mpg.de

[2] Faculty of Mathematics and Computer Science, University of Leipzig, PF 100920, 04009 Leipzig, Germany

[3] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Academic Editors: Christoph Salge, Georg Martius, Keyan Ghazi-Zahedi, and Daniel Polani

Received: 28 February 2015 / Accepted: 8 April 2015 / Published: 21 April 2015

Abstract: Interdependencies of stochastically interacting units are usually quantified by the Kullback-Leibler divergence of a stationary joint probability distribution on the set of all configurations from the corresponding factorized distribution. This is a spatial approach which does not describe the intrinsically temporal aspects of interaction. In the present paper, the setting is extended to a dynamical version where temporal interdependencies are also captured by using information geometry of Markov chain manifolds.

Keywords: stochastic interaction; complexity; information geometry; Kullback-Leibler divergence; separability; Markov chains; random fields

---

1. Preface: Information Integration and Complexity

Since the publication of Shannon's pioneering work in 1948 [1], it has been hypothesized that his information theory provides means for understanding information processing and learning in the brain. Already in the 1950s, the principle of redundancy reduction has been proposed independently by Attneave [2] and Barlow [3]. In 1981, Laughlin has provided some experimental evidence for the redundancy reduction principle in terms of the maximization of the output entropy of large monopolar cells of the fly's compound eye [4]. As only deterministic response functions have been considered, this principle turns out to be equivalent to the mutual information maximization between the input and the output. Later, Linsker [5] has demonstrated that the maximization of mutual information in a layered feed-forward network leads to feature detectors that are similar to those observed by Hubel and Wiesel

in the visual system of the cat and the monkey [6,7]. He coined his information-theoretic principle of learning the *infomax principle*.

The idea that an information-theoretic principle, such as the infomax principle, governs learning processes of neuronal systems has attracted many researchers. A highly recognized contribution in this regard is the work by Bell and Sejnowski [8] which applies the infomax principle to the source separation problem. An exhaustive review of all relevant contributions to that field is not within the scope of this short discussion. I shall focus on approaches that aim at relating such information based principles to the overall complexity of the system. In particular, I shall concentrate on the theory of information integration and complexity, initially proposed by Tononi, Sporns, and Edelman [9], and further developed and analyzed in a series of papers [10–15]. I shall compare this line of research with my own information-geometric approach to complexity, initially proposed in my manuscript [16], entitled *Information Geometry on Complexity and Stochastic Interaction*, which led to various lines of research that I am going to outline below. This manuscript constitutes the main body of the present paper, starting with Section 2. It quantifies complexity as the extent to which the whole is more than the sum of its parts using information geometry [17]. Thereby, it extends the notion of multi-information [18,19], also called information integration in [9], to the setting of discrete time stochastic processes, in particular Markov chains. This article was originally accepted for publication in IEEE Transactions on Information Theory, subject to minor revision. However, by the end of the unusually long reviewing process I had come to the conclusion that my geometric approach has to be further improved in order to address important aspects of complexity (I shall be more concrete on that). Recent developments, on the other hand, suggest that this work is of relevance in the context of information integration already in its present form [12–15,20,21]. Therefore, it should be useful to provide it together with a discussion of its strengths and shortcomings, thereby relating it to similar work that has been developed since its first publication.

Let us first consider the so-called *multi-information* [18,19] of a random vector $X = (X_v)_{v \in V}$, taking values in a finite set:

$$I(X) := \sum_{v \in V} H(X_v) - H(X)\,, \tag{1}$$

where H denotes the Shannon entropy (we assume V to be a non-empty and finite set). The multi-information vanishes if and only if the variables X_v, $v \in V$, are stochastically independent. In their original paper [9], Tononi, Sporns, and Edelman call this quantity *integration*. Following their intuition, however, the notion of integration should rather refer to a dynamical process, the process of integration, which is causal in nature. In later works, the dynamical aspects have been more explicitly addressed in terms of a causal version of mutual information, leading to improved notions of *effective information* and *information integration*, denoted by Φ [10,11]. In fact, most formulated information-theoretic principles are, in some way or another, based on (conditional) mutual information. This directly fits into Shannon's classical sender-receiver picture [1], where the mutual information has been used in order to quantify the capacity of a communication channel. At first sight, this picture suggests to treat only feed-forward networks, in which information is transmitted from one layer to the next, as in the context of Linsker's infomax principle. In order to overcome this apparent restriction, however, we can simply unfold the dynamics in time and consider corresponding temporal information

flow measures, which allows us to treat also recurrent networks. In what follows, I am going to explain this idea in more detail, thereby providing a motivation of the quantities that are derived in Section 2 in terms of information geometry.

We consider again a non-empty and finite set V of nodes and assume that each $v \in V$ receives signals from a set of nodes which we call *parents* of v and denote by $pa(v)$. Based on the received signals, the node v updates its state according to a Markov kernel $K^{(v)}$, the mechanism of v, which quantifies the conditional probability of its new state ω'_v given the current state $\omega_{pa(v)}$ of its parents. If $v \in pa(v)$, this update will involve also ω_v for generating the new state ω'_v . How much information is involved from "outside", that is from $\partial(v) := pa(v) \setminus v$, in addition to the information given by ω_v? We can define the *local information flow* from this set as

$$IF(X_{\partial(v)} \to X'_v) \;:=\; H(X'_v \,|\, X_v) - H(X'_v \,|\, X_v, X_{\partial(v)}) \;=\; MI(X'_v; X_{\partial(v)} \,|\, X_v)\,, \tag{2}$$

where MI stands for the (conditional) mutual information. Note that this is the uncertainty reduction that the node v gains through the knowledge of its parents' state, in addition to its own state. Now let us define the total information flow in the network. In order to do so, we have to consider the overall transition kernel. Because the nodes update their states in parallel, the global transition kernel is given as

$$K(\omega' \,|\, \omega) \;=\; \prod_{v \in V} K^{(v)}(\omega'_v \,|\, \omega_{pa(v)})\,. \tag{3}$$

In order to quantify the *total information flow* in the network, we simply add all the local information flows, defined by Equation (2), and obtain

$$IF(X \to X') \;:=\; \sum_{v \in V} IF(X_{\partial(v)} \to X_v)\,. \tag{4}$$

It is easy to see that the total information flow vanishes whenever the global transition kernel has the following structure which encodes the dynamics of isolated non-communicating nodes:

$$K(\omega' \,|\, \omega) \;=\; \prod_{v \in V} K^{(v)}(\omega'_v \,|\, \omega_v)\,. \tag{5}$$

Referring to these kernels as being *split*, we are now ready to give our network information flow measure, defined by Equation (4), a geometric interpretation. If K has the structure Equation (3) then

$$IF(X \to X') \;=\; \sum_{v \in V} H(X'_v \,|\, X_v) - H(X' \,|\, X) \tag{6}$$

$$\;=\; \min_{K' \text{ split}} D_p(K \,\|\, K')\,. \tag{7}$$

Here, $D_p(K \,\|\, K')$ is a measure of "distance", in terms of the Kullback-Leibler divergence, between K and K' with respect to the distribution p (see definition by Equation (23)). The expression on the right-hand side of Equation (6) can be considered as an extension of the multi-information (1) to the temporal domain. The second equality, Equation (7), gives the total information flow in the network a geometric interpretation as the distance of the global dynamics K from the set of split dynamics. Stated differently, the total information flow can be seen as the extent to which the whole transition $X \to X'$ is more than the sum of its individual transitions $X_v \to X'_v$, $v \in V$. Note, however, that

Equation (6) follows from the additional structure (3) which implies $H(X' \mid X) = \sum_{v \in V} H(X'_v \mid X)$. This structure encodes the consistency of the dynamics with the network. Equation (7), on the other hand, holds for any transition kernel K. Therefore, without reference to a particular network, the distance $\min_{K' \text{ split}} D_p(K \,\|\, K')$ can be considered as a complexity measure for any transition $X \to X'$, which we denote by $C^{(1)}(X \to X')$. The information-geometric derivation of $C^{(1)}(X \to X')$ is given in Section 2.4.1. Restricted to kernels that are consistent with a network, the complexity $C^{(1)}(X \to X')$ reduces to the total information flow in the network (see Proposition 2 (iv)).

In order to consider the maximization of the complexity measure $C^{(1)}(X \to X')$ as a valid information-theoretic principle of learning in neuronal systems, I analyzed the natural gradient field on the manifold of kernels that have the structure given by Equation (3) (see [17,22] for the natural gradient method within information geometry). In [23] I proved the consistency of this gradient in the sense that it is completely local: If every node v maximizes its own local information flow, defined by Equation (2), in terms of the natural gradient, then this will be the best way, again with respect to the natural gradient, to maximize the complexity of the whole system. This suggests that the infomax principle by Linsker and also Laughlin's ansatz, applied locally to recurrent networks, will actually lead to the maximization of the overall complexity. We used geometric methods to study the maximizers of this complexity analytically [24,25]. We have shown that they are almost deterministic, which has quite interesting implications, for instance for the design of learning systems that are parametrized in a way that allows them to maximize their complexity [26] (see also [27] for an overview of geometric methods for systems design). Furthermore, evidence has been provided in [25] that the maximization of $C^{(1)}(X \to X')$ is achieved in terms of a rule that mimics the spike-timing-dependent plasticity of neurons in the context of discrete time. Together with Wennekers, we have studied complexity maximization as first principle of learning in neural networks also in [28–33].

Even though I implicitly assumed that a natural notion of information flow has to reflect the causal interactions of the nodes, I should point out that the above definition of information flow has a shortcoming in this regard. If X_v and $X_{\partial(v)}$ contain the same information, due to a strong stochastic dependence, then the conditional mutual information in Equation (2) will vanish, even though there might be a strong causal effect of $\partial(v)$ on v. Thus, correlation among various potential causes can hide the actual causal information flow. The information flow measure of Equation (2) is one instance of the so-called transfer entropy [34] which is used within the context of Granger causality and has, as a conditional mutual information, the mentioned shortcoming also in more general settings (see a more detailed discussion in [35]). In order to overcome these limitations of the (conditional) mutual information, in a series of papers [35–39] we have proposed the use of information theory in combination with Pearl's theory of causation [40]. Our approach has been discussed in [41] where a variant of our notion of node exclusion, introduced in [36], has been utilized for an alternative definition. This definition, however, is restricted to direct causal effects and does not capture, in contrast to [35], mediated causal effects.

Let us now draw a parallel to causality issues of the complexity measure introduced in the original work [9], which we refer to as *TSE-complexity*. In order to do so, consider the following representation of the original TSE-complexity as weighted sum of mutual informations:

$$C_{TSE}(X) \;:=\; \sum_{A\subseteq V} \alpha_A \, MI(X_A; X_{V\setminus A})\,, \tag{8}$$

where $\alpha_A = \frac{k}{N\binom{N}{k}}$. Interpreting the mutual information between A and its complement $V \setminus A$ in this sum as an information flow is clearly misleading. These terms are completely associational and neglect the causal nature of information flow. In [10,11], Tononi and Sporns avoid such inconsistencies by injecting noise (maximum entropy distribution) into A and then measuring the effect in $V \setminus A$. They use the corresponding interventional mutual information in order to define *effective information*. Note that, although their notion of noise injection is conceptually similar to the notion of intervention proposed by Pearl, they formalize it differently. However, the idea of considering a post-interventional mutual information is similar to the one formalized in [35,36] using Pearl's interventional calculus.

Clearly, the measure $C^{(1)}(X \to X')$ does not account for all aspects of the system's complexity. One obvious reason for that can be seen by comparison with the multi-information, defined by Equation (1), which also captures some aspects of complexity in the sense that it quantifies the extent to which the whole is more than the sum of its elements (parts of size one). On the other hand, it attains its (globally) maximal value, if and only if the nodes are completely correlated. Such systems, in particular completely synchronized systems, are generally not considered to be complex. Furthermore, it turns out that these maximizers are determined by the marginals of size two [42]. Stated differently, the maximization of the extent to which the whole is more than the sum of its parts of size one leads to systems that are not more than the sum of their parts of size two (see for a more detailed discussion [43,44]). Therefore, the multi-information does not capture the complexity of a distribution at all levels. The measure $C^{(1)}(X \to X')$ has the same shortcoming as the multi-information. In order to study different levels of complexity, one can consider coarse-grainings of the system at different scales in terms of corresponding partitions $\Pi = \{S_1, \ldots, S_n\}$ of V. Given such a partition, we can define the information flows among its atoms S_i as we already did for the individual elements v of V. For each S_i, we denote the set of nodes that provide information to S_i from outside by $\partial(S_i) := \bigcup_{v\in S_i} (pa(v) \setminus S_i)$. We quantify the information flow into S_i as in Equation (2):

$$IF(X_{\partial(S_i)} \to X_{S_i}) \;:=\; H(X'_{S_i} \,|\, X_{S_i}) - H(X_{S'_i} \,|\, X_{S_i}, X_{\partial(S_i)}) \;=\; MI(X'_{S_i}; X_{\partial(S_i)} \,|\, X_{S_i})\,. \tag{9}$$

For a transition that satisfies Equation (3), the total information flow among the parts S_i is then given by

$$IF(X \to X' \,|\, \Pi) \;:=\; \sum_{i=1}^{n} IF(X_{\partial(S_i)} \to X_{S_i})\,. \tag{10}$$

We can now define the Π-*complexity* of a general transition, as we already did for the complete partition:

$$C(X \to X' \,|\, \Pi) \;:=\; \sum_{i=1}^{n} H(X'_{S_i} \,|\, X_{S_i}) - H(X' \,|\, X)\,. \tag{11}$$

Obviously, the Π-*complexity* coincides with the information flow $IF(X \to X' \,|\, \Pi)$ in the case where the transition kernel is compatible with the network. The information-geometric derivation of $C(X \to X' \,|\, \Pi)$ is given in Section 2.4.1. In the early work [10,11], a similar approach has been proposed where only bipartitions have been considered. Later, an extension to arbitrary partitions has

been proposed by Balduzzi and Tononi [12,13] where the complexity defined by Equation (11) appears as measure of effective information. Note, however, that there are important differences. First, the proposed measure by Tononi and his coworkers is reversed in time, so that their quantity is given by Equation (11) where X and X' have exchanged roles. This time-reversal of the effective information is motivated by its intended role as a measure relevant to conscious experience. This does not make any difference in the case where a stationary distribution is chosen as input distribution. However, in order to be consistent with causal aspects of conscious experience, the authors choose a uniform input distribution, which models the least informative prior about the input.

Note that there is also a closely related measure, referred to as *synergistic information* in the works [15,45]:

$$SI(X \to X' \,|\, \Pi) \;:=\; MI(X';X) - \sum_{i=1}^{n} MI(X'_{S_i}; X_{S_i}) \tag{12}$$

$$\;=\; C(X \to X' \,|\, \Pi) - I(X_{S_1}, \ldots, X_{S_n})\,. \tag{13}$$

The last equation directly follows from Proposition 1 (iii) (see the derivation of Equation (29)). Interpreting the mutual informations as (one-step) predictive information [46–48], the synergistic information quantifies the extent to which the predictive information of the whole system exceeds the sum of predictive informations of the elements.

Now, having for each partition of the system the corresponding Π-complexity of Equation (11), how should one choose among all these complexities the right one? Following the proposal made in [10–13], one should identify the partition (or bipartition) that has the smallest, appropriately normalized, Π-complexity. Although the overall complexity is not explicitly defined in these works, the notion of *information integration*, denoted by Φ, seems to directly correspond to it. This is confirmed by the fact that information integration is used for the identification of so-called *complexes* in the system. Loosely speaking, these are defined to be subsets S of V with maximal information integration $\Phi(S)$. This suggests that the authors equate information integration with complexity. In a further refinement [12,13] of the information integration concept, this is made even more explicit. In [13], Tononi writes: "In short, integrated information captures the information generated by causal interactions in the whole, over and above the information generated by the parts."

Defining the overall complexity simply as the minimal one, with respect to all partitions, will ensure that a complex system has a considerably high complexity at *all* levels. I refer to this choice as the *weakest link approach*. This is not the only approch to obtain an overall complexty measure from individual ones defined for various levels. In order to give an instructive example for an alternative approach, let us highlight another representation of the TSE-complexity. Instead of the atoms of a partition, this time we consider the subsets of V with a given size $k \in \{1, \ldots, N\}$ and define the following quantity:

$$C^{(k)}(X) \;:=\; \frac{N}{k\binom{N}{k}} \sum_{\substack{A \subseteq V \\ |A|=k}} H(X_A) - H(X)\,. \tag{14}$$

Let us compare this quantity with the multi-information of Equation (1). For $k = 1$, they are identical. While the multi-information quantifies the extent to which the whole is more than the sum of its elements (subsets of size one), its generalization $C^{(k)}(X)$ can be interpreted as the extent to which the whole is

more than the sum of its parts of size k. Now, defining the overall complexity as the minimal $C^{(k)}(X)$ would correspond to the weakest link approach which I discussed above in the context of partitions. A complex system would then have considerably high complexity $C^{(k)}(X)$ at all levels k. However, the TSE-complexity is not constructed according to the weakest link approach, but can be written as a weighted sum of the terms $C^{(k)}(X)$:

$$C_{TSE}(X) \;=\; \sum_{k=1}^{N} \alpha(k)\, C^{(k)}(X)\,, \tag{15}$$

where $\alpha(k) = \frac{k}{N}$. The right choice of the weights is important here. I refer to this approach as the *average approach*. Clearly, one can interpolate between the weakest link approach and the average approach using the standard interpolation between the L^{∞}-norm (maximum) and the L^{1}-norm (average) in terms of the L^{p}-norms, $p \geq 1$. However, L^{p}-norms appear somewhat unnatural for entropic quantities.

The TSE-complexity has also an information-geometric counterpart which has been developed in a series of papers [43,44,49,50]. It is instructive to consider this geometric reformulation of the TSE-complexity. For a distribution p, let $p^{(k)}$ be the maximum-entropy estimation of p with fixed marginals of order k. In particular, $p^{(N)} = p$, and $p^{(1)}$ is the product of the marginals p_v, $v \in V$, of order one. In some sense, $p^{(k)}$ encodes the structure of p that is contained only in the parts of size k. The deviation of p from $p^{(k)}$ therefore corresponds to $C^{(k)}(X)$, as defined in Equation (14). This correspondence can be made more explicit by writing this deviation in terms of a difference of entropies:

$$D(p \,\|\, p^{(k)}) \;=\; H_{p^{(k)}}(X) - H_p(X)\,, \tag{16}$$

where D denotes the Kullback-Leibler divergence. If we compare the Equations (16) and (14), then we see that $\frac{N}{k\binom{N}{k}} \sum_{\substack{A \subseteq V \\ |A|=k}} H(X_A)$ corresponds to $H_{p^{(k)}}(X)$. Indeed, both terms quantify the entropy that is contained in the marginals of order k. From the information-geometric point of view, however, the second term appears more natural. The first term seems to count marginal entropies multiple times so that we can expect that this mean value is larger than $H_{p^{(k)}}(X)$. In [43], we have shown that this is indeed true, which implies

$$D(p \,\|\, p^{(k)}) \;\leq\; C^{(k)}(X)\,. \tag{17}$$

If we replace the $C^{(k)}(X)$ in the definition (15) of the TSE-complexity by $D(p \,\|\, p^{(k)})$, then we obtain with the Pythagorean theorem of information geometry the following quantity:

$$I^{\beta}(X) \;:=\; \sum_{k=1}^{N-1} \beta(k)\, D(p^{(k+1)} \,\|\, p^{(k)})\,, \tag{18}$$

where $\beta(k) = \frac{k(k+1)}{2}$. Let us compare this with the multi-information. Following [18], we can decompose the multi-information as

$$I(X) \;=\; D(p \,\|\, p^{(1)}) \;=\; \sum_{k=1}^{N-1} D(p^{(k+1)} \,\|\, p^{(k)})\,. \tag{19}$$

I already mentioned that high multi-information is achieved for strongly correlated systems, which implies that the global maximizers can be generated by systems that only have pairwise interactions [42],

that is $p = p^{(2)}$. It follows that in the above decompsition of Equation (19), only the first term $D(p^{(2)} \,\|\, p^{(1)})$ is positive while all the other terms vanish for maximizers of the multi-information. This suggests that the multi-information does not weight all contributions $D(p^{(k+1)} \,\|\, p^{(k)})$ to the stochastic dependence in a way that would qualify it as a complexity measure. The measure defined by Equation (18), which I see as an information-geometric counterpart of the TSE-complexity, weights the higher-order contributions $D(p^{(k+1)} \,\|\, p^{(k)})$, $k \geq 2$, more strongly. In this geometric picture, we can interpret the TSE-complexity as a rescaling of the multi-information in such a way that its maximization will emphasize not only pairwise interactions.

Concluding this preface, I compared two lines of research, the one pursued by Tononi and coworkers on information integration, and my own information-geometric research on complexity. The fact that both research lines independently identified closely related core concepts of complexity confirms that these concepts are quite natural. The comparison of the involved ideas suggests the following intuitive definition of complexity: *The complexity of a system is the extent to which the whole is more than the sum of its parts at all system levels.* I argue that information geometry provides natural methods for casting this intuitive definition into a formal and quantitative theory of complexity. My paper [16], included here as Section 2, exemplifies this way of thinking about complexity. It is presented with only minor changes compared to its initial publication, except that the original reference list is replaced by the largely extended up-to-date list of references. This implies repetitions of a few standard definitions which I already used in this preface.

2. "Information Geometry on Complexity and Stochastic Interaction", Reference [16]

2.1. Introduction

"The whole is more than the sum of its elementary parts." This statement characterizes the present approach to *complexity*. Let us put it in a more formal setting. Assume that we have a system consisting of elementary units $v \in V$. With each non-empty subsystem $S \subset V$ we associate a set $\mathcal{O}_S$ of objects that can be generated by S. Examples for such objects are (deterministic) dynamical systems, stochastic processes, and probability distributions. Furthermore, we assume that there is a "composition" map $\otimes : \prod_{v \in V} \mathcal{O}_{\{v\}} \hookrightarrow \mathcal{O}_V$ that defines how to put objects of the individual units together in order to describe a global object without any interrelations. The image of $\otimes$ consists of the *split* global objects which are completely characterized by the individual ones and therefore represent the absence of complexity. In order to quantify complexity, assume that there is given a function $D : (x, y) \mapsto D(x \,\|\, y)$, that measures the divergence of global objects $x, y \in \mathcal{O}_V$. We define the complexity of $x \in \mathcal{O}_V$ to be the divergence from being split:

$$\textit{Complexity}\,(x) \;:=\; \inf_{y \text{ split}} D(x \,\|\, y)\,. \tag{20}$$

Of course, this approach is very general, and there are many ways to define complexity following this concept. Is there a canonical way? At least, within the probabilistic setting, *information geometry* [17,51] provides a very convincing framework for this. In the context of random fields, it leads to a measure for "spatial" interdependencies: Given state sets Ω_v, $v \in V$, we define the set $\mathcal{O}_S$ of objects that are generated by a subsystem $S \subset V$ to be the probability distributions

on the product set $\prod_{v\in S}\Omega_v$. A family of individual probability distributions $p^{(v)}$ on Ω_v can be considered as a distribution on the whole configuration set $\prod_{v\in V}\Omega_v$ by identifying it with the product $\otimes_{v\in V}\, p^{(v)} \in \mathcal{O}_V$. In order to define the complexity of a distribution $p \in \mathcal{O}_V$ on the whole system, according to Equation (20) we have to choose a divergence function. A canonical choice for D is given by the *Kullback-Leibler divergence* [52,53]:

$$\textit{Complexity}\,(p) \;:=\; I(p) \;:=\; \inf_{p^{(v)}\in\mathcal{O}_v,\; v\in V} D\big(p \,\|\, \otimes_{v\in V}\; p^{(v)}\big)\,. \tag{21}$$

It is well known that $I(p)$ quantifies spatial interdependencies [18]. It vanishes exactly when the units are stochastically independent with respect to p. Such split distributions are called *factorizable* in this context. In Figure 1, the example of two binary units with the state sets $\{0,1\}$ is illustrated.

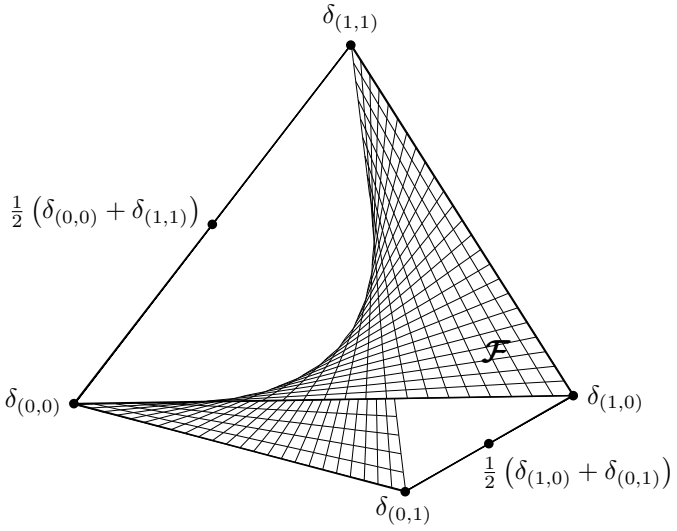

Figure 1. $\mathcal{F}$ denotes the set of factorizable distributions on $\{0,1\}\times\{0,1\}$.

The distributions with maximal interdependence (complexity) are given by

$$\frac{1}{2}\left(\delta_{(0,0)}+\delta_{(1,1)}\right) \qquad \text{and} \qquad \frac{1}{2}\left(\delta_{(1,0)}+\delta_{(0,1)}\right).$$

Spatial interdependence has been studied by Amari [18] and Ay [23,55] from the information-geometric point of view, where it is referred to as (*stochastic*) *interaction* and discussed in view of neural networks. The aim of the present paper is to use the concept of complexity that is formalized by Equation (20) in order to extend spatial interdependence to a dynamical notion of interaction, where the evolution in time is taken into account. Therefore, the term "stochastic interaction" is mainly used in the context of spatio-temporal interdependence.

The present paper is organized as follows. After a brief introduction into the information-geometric description of finite probability spaces in Section 2.2, the general notion of separability is introduced for Markovian transition kernels, and information geometry is used for quantifying non-separability as divergence from separability (Section 2.3). In Section 2.4, the presented theoretical framework is used to derive a dynamical version of the definition in Equation (21), where spatio-temporal interdependencies are quantified and referred to as *stochastic interaction*. This is illustrated by some simple but instructive examples.

2.2. Preliminaries on Finite Information Geometry

In the following, Ω denotes a non-empty and finite set. The vector space $\mathbb{R}^\Omega$ of all functions $\Omega \to \mathbb{R}$ carries the natural topology, and we consider subsets as topological subspaces. The set of all probability distributions on Ω is given by

$$\bar{\mathcal{P}}(\Omega) \;:=\; \left\{ p = (p(\omega))_{\omega\in\Omega} \in \mathbb{R}^\Omega \;:\; p(\omega) \geq 0 \;\text{ for all }\; \omega \in \Omega, \quad \sum_{\omega\in\Omega} p(\omega) = 1 \right\} .$$

Following the information-geometric description of finite probability spaces, its interior $\mathcal{P}(\Omega)$ can be considered as a differentiable submanifold of $\mathbb{R}^\Omega$ with dimension $|\Omega|-1$ and the basis-point independent tangent space

$$\mathrm{T}(\Omega) := \left\{ x \in \mathbb{R}^\Omega \;:\; \sum_{\omega\in\Omega} x(\omega) = 0 \right\} .$$

(If one considers $\mathcal{P}(\Omega)$ as an "abstract" differentiable manifold, there are many ways to represent it as a submanifold of $\mathbb{R}^\Omega$. In information geometry, the natural embedding presented here is called (-1)-respectively (m)-*representation*)

With the *Fisher metric* $\langle\cdot,\cdot\rangle_p : \mathrm{T}(\Omega)\times\mathrm{T}(\Omega) \to \mathbb{R}$ in $p \in \mathcal{P}(\Omega)$ defined by

$$(x,y) \;\;\mapsto\;\; \langle x,y\rangle_p := \sum_{\omega\in\Omega} \frac{1}{p(\omega)}\, x(\omega) y(\omega) \,,$$

$\mathcal{P}(\Omega)$ becomes a Riemannian manifold [56] (In mathematical biology this metric is also known as *Shahshahani metric* [57]). The most important additional structure studied in information geometry is given by a pair of dual affine connections on the manifold. Application of such a dual structure to the present situation leads to the notion of (-1)- and $(+1)$-geodesics: Each two points $p, q \in \mathcal{P}(\Omega)$ can be connected by the geodesics $\gamma^{(\alpha)} = \left(\gamma_\omega^{(\alpha)}\right)_{\omega\in\Omega} :\; [0,1] \to \mathcal{P}(\Omega)$, $\alpha \in \{-1,+1\}$, with

$$\gamma_\omega^{(-1)}(t) \;:=\; (1-t)\, p(\omega) + t\, q(\omega) \quad\text{and}\quad \gamma_\omega^{(+1)}(t) \;:=\; r(t)\; p(\omega)^{1-t}\, q(\omega)^t \,.$$

Here, $r(t)$ denotes the normalization factor.

A submanifold $\mathcal{E}$ of $\mathcal{P}(\Omega)$ is called an *exponential family* if there exist a point $p_0 \in \mathcal{P}(\Omega)$ and vectors $v_1, \ldots, v_d \in \mathbb{R}^\Omega$, such that it can be expressed as the image of the map $\mathbb{R}^d \to \mathcal{P}(\Omega)$, $\theta = (\theta_1, \ldots, \theta_d) \mapsto p_\theta$, with

$$p_\theta(\omega) \;:=\; \frac{p_0(\omega)\, \exp\left(\sum_{i=1}^d \theta_i\, v_i(\omega)\right)}{\sum_{\omega'\in\Omega} p_0(\omega')\, \exp\left(\sum_{i=1}^d \theta_i\, v_i(\omega')\right)} \,. \tag{22}$$

Let p be a probability distribution in $\mathcal{P}(\Omega)$. An element $p' \in \mathcal{E}$ is called (-1)-*projection* of p onto $\mathcal{E}$ iff the (-1)-geodesic connecting p and p' intersects $\mathcal{E}$ orthogonally with respect to the Fisher metric. Such a point p' is unique ([51], Theorem 3.9, p. 91) and can be characterized by the *Kullback-Leibler divergence* [52,53] (This is a special case of Csiszár's f-*divergence* [54])

$$D : \mathcal{P}(\Omega)\times\mathcal{P}(\Omega) \to \mathbb{R}_+, \qquad (p,q) \;\;\mapsto\;\; D(p\,\|\,q) \;:=\; \sum_{\omega\in\Omega} p(\omega)\, \ln\frac{p(\omega)}{q(\omega)} \,.$$

We define the distance $D(\cdot \,\|\, \mathcal{E}) : \mathcal{P}(\Omega) \to \mathbb{R}_+$ from $\mathcal{E}$ by

$$p \;\mapsto\; D(p \,\|\, \mathcal{E}) \;:=\; \inf_{q \in \mathcal{E}} D(p \,\|\, q)\,.$$

It is well known that a point $p' \in \mathcal{E}$ is the (-1)-projection of p onto $\mathcal{E}$ if and only if it satisfies the minimizing property $D(p \,\|\, \mathcal{E}) = D(p \,\|\, p')$ ([51], Theorem 3.8, p. 90; [17], Corollary 3.9, p. 63).

In the present paper, the set of states is given by the Cartesian product of individual state sets Ω_v, $v \in V$, where V denotes the set of *units*. In the following, the unit set and the corresponding state sets are assumed to be non-empty and finite. For a subsystem $S \subset V$, $\Omega_S := \prod_{v \in S} \Omega_v$ denotes the set of all configurations on S. The elements of $\bar{\mathcal{P}}(\Omega_S)$ are the *random fields* on S. One has the natural restriction $X_S : \Omega_V \to \Omega_S$, $\omega = (\omega_v)_{v \in V} \mapsto \omega_S := (\omega_v)_{v \in S}$, which induces the projection $\bar{\mathcal{P}}(\Omega_V) \to \bar{\mathcal{P}}(\Omega_S)$, $p \mapsto p_S$, where p_S denotes the image measure of p under the variable X_S. If the subsystem S consists of exactly one unit v, we write p_v instead of $p_{\{v\}}$.

The following example, which allows us to put the definition of Equation (21) into the information-geometric setting, represents the main motivation for the present approach to stochastic interaction. It will be generalized in Section 2.4.

Example 1 (FACTORIZABLE DISTRIBUTIONS AND SPATIAL INTERDEPENDENCE)**.** Let V be a finite set of units and Ω_v, $v \in V$, corresponding state sets. Consider the *tensorial map*

$$\prod_{v \in V} \mathcal{P}(\Omega_v) \;\hookrightarrow\; \mathcal{P}(\Omega_V), \qquad (p^{(v)})_{v \in V} \;\mapsto\; \otimes_{v \in V} p^{(v)}\,,$$

with

$$\left(\otimes_{v \in V} p^{(v)}\right)(\omega) \;:=\; \prod_{v \in V} p^{(v)}(\omega_v)\,.$$

The image $\mathcal{F} := \mathcal{F}(\Omega_V) := \left\{\otimes_{v \in V} p^{(v)} \,:\, p^{(v)} \in \mathcal{P}(\Omega_v),\;\; v \in V\right\}$ of this map, which consists of all factorizable and strictly positive probability distributions, is an exponential family in $\mathcal{P}(\Omega_V)$ with $\dim \mathcal{F} = \sum_{v \in V}(|\Omega_v| - 1)$. For the particular case of binary units, that is $|\Omega_v| = 2$ for all v, the dimension of $\mathcal{F}$ is equal to the number $|V|$ of units. The following statement is well known [18]: The (-1)-projection of a distribution $p \in \mathcal{P}(\Omega_V)$ on $\mathcal{F}$ is given by $\otimes_{v \in V} p_v$ (the p_v, $v \in V$, are the marginal distributions), and one has the representation

$$I(p) \;=\; D(p \,\|\, \mathcal{F}) \;=\; \sum_{v \in V} H(p_v) - H(p)\,,$$

where H denotes the *Shannon entropy* [1]. As stated in the introduction, $I(p)$ is a measure for the spatial interdependencies of the units. It vanishes exactly when the units are stochastically independent.

Before extending the spatial notion of interaction to a dynamical one, in Section 2.3 we consider the more general concept of separability of transition kernels.

2.3. Quantifying Non-Separability

2.3.1. Manifolds of Separable Transition Kernels

Consider a finite set V of units, corresponding state sets Ω_v, $v \in V$, and two subsets $A, B \subset V$ with $B \neq \emptyset$. A function

$$K : \Omega_A \times \Omega_B \to [0,1], \qquad (\omega, \omega') \mapsto K(\omega' \,|\, \omega)\,,$$

is called *Markovian transition kernel* if $K(\cdot \,|\, \omega) \in \bar{\mathcal{P}}(\Omega_B)$ for all $\omega \in \Omega_A$, that is

$$\sum_{\omega' \in \Omega_B} K(\omega' \,|\, \omega) \;=\; 1, \qquad \text{for all} \quad \omega \in \Omega_A\,.$$

The set of all such kernels is denoted by $\bar{\mathcal{K}}(\Omega_B \,|\, \Omega_A)$. We write $\mathcal{K}(\Omega_B \,|\, \Omega_A)$ for its interior and $\bar{\mathcal{K}}(\Omega_A)$ respectively $\mathcal{K}(\Omega_A)$ as abbreviation in the case $A = B$. If $A = \emptyset$, then Ω_A consists of exactly one element, namely the empty configuration ϵ. In that case, $\bar{\mathcal{K}}(\Omega_B \,|\, \Omega_\emptyset) = \bar{\mathcal{K}}(\Omega_B \,|\, \epsilon)$ can naturally be identified with $\bar{\mathcal{P}}(\Omega_B)$ by $p(\omega) := K(\omega \,|\, \epsilon)$, $\omega \in \Omega_B$.

Given a probability distribution $p \in \bar{\mathcal{P}}(\Omega_A)$ and a transition kernel $K \in \bar{\mathcal{K}}(\Omega_B \,|\, \Omega_A)$, the *conditional entropy* for (p, K) is defined as

$$H(p, K) \;:=\; \sum_{\omega \in \Omega_A} p(\omega)\, H\big(K(\cdot \,|\, \omega)\big)\,.$$

For two random variables X, Y with $\mathrm{Prob}\{X = \omega\} = p(\omega)$ for all $\omega \in \Omega_A$, and $\mathrm{Prob}\{Y = \omega' \,|\, X = \omega\} = K(\omega' \,|\, \omega)$ for all $\omega \in \Omega_A$ with $p(\omega) > 0$ and all $\omega' \in \Omega_B$, we set $H(Y \,|\, X) := H(p, K)$.

In the present paper, the set $\bar{\mathcal{K}}(\Omega_V)$ is interpreted as a model for the dynamics of interacting units, and the information flow associated with this dynamics is studied in Section 2.4. In the present section, we introduce a general notion of separability of transition kernels in order to capture all examples that are discussed in the paper in a unified way.

Consider a family $\mathscr{S} := \{(A_1, B_1), (A_2, B_2), \ldots, (A_n, B_n)\}$ where the A_i and B_i are subsets of V. We assume that $\{B_1, \ldots, B_n\}$ is a partition of V, that is $B_i \neq \emptyset$ for all i, $B_i \cap B_j = \emptyset$ for all $i \neq j$, and $V = B_1 \uplus \cdots \uplus B_n$. Now consider the corresponding *tensorial map*

$$\otimes_{\mathscr{S}} : \prod_{(A,B) \in \mathscr{S}} \mathcal{K}(\Omega_B \,|\, \Omega_A) \;\hookrightarrow\; \mathcal{K}(\Omega_V), \quad \big(K_B^A\big)_{(A,B) \in \mathscr{S}} \;\mapsto\; \otimes_{(A,B) \in \mathscr{S}} K_B^A\,,$$

with

$$\big(\otimes_{(A,B) \in \mathscr{S}} K_B^A\big)\,(\omega' \,|\, \omega) \;:=\; \prod_{(A,B) \in \mathscr{S}} K_B^A(\omega'_B \,|\, \omega_A), \qquad \text{for all} \quad \omega, \omega' \in \Omega_V\,.$$

The image $\mathcal{K}_{\mathscr{S}}(\Omega_V)$ of $\otimes_{\mathscr{S}}$ is a submanifold of $\mathcal{K}(\Omega_V)$ with

$$\dim \mathcal{K}_{\mathscr{S}}(\Omega_V) \;=\; \sum_{(A,B) \in \mathscr{S}} |\Omega_A|\big(|\Omega_B| - 1\big)\,.$$

Its elements are the *separable* transition kernels with respect to $\mathscr{S}$.

Here are the most important examples:

Examples and Definitions 1.
(1) If we set $\mathscr{S} := \{(V,V)\}$, the tensorial map is nothing but the identity $\mathcal{K}(\Omega_V) \to \mathcal{K}(\Omega_V)$, and therefore one has $\mathcal{K}_{\mathscr{S}}(\Omega_V) = \mathcal{K}(\Omega_V)$.

(2) Consider the case where no temporal information is transmitted but all spatial information: $\mathscr{S} := ind := \{(\emptyset, V)\}$. In that case the tensorial map $\otimes_{\mathscr{S}}$ reduces to the natural embedding

$$\mathcal{K}(\Omega_V \,|\, \Omega_\emptyset) = \mathcal{P}(\Omega_V) \hookrightarrow \mathcal{K}(\Omega_V)$$

which assigns to each probability distribution p the kernel

$$K(\omega' \,|\, \omega) := p(\omega'), \qquad \omega, \omega' \in \Omega_V \,.$$

Therefore, we write $\mathcal{K}_{ind}(\Omega_V) = \mathcal{P}(\Omega_V)$.

(3) In addition to the splitting in time which is described in example (2), consider also a complete splitting in space: $\mathscr{S} := fac := \{(\emptyset, \{v\}) \,:\, v \in V\}$. Then we recover the tensorial map of Example 1. Thus, $\mathcal{K}_{fac}(\Omega_V)$ can be identified with $\mathcal{F}(\Omega_V)$.

(4) To model the important class of *parallel information processing*, we set $\mathscr{S} := par := \{(V, \{v\}) \,:\, v \in V\}$. Here, each unit "computes" its new state on the basis of all current states according to a kernel $K^{(v)} \in \mathcal{K}(\Omega_v \,|\, \Omega_V)$. The transition from a configuration $\omega = (\omega_v)_{v \in V}$ of the whole system to a new configuration $\omega' = (\omega'_v)_{v \in V}$ is done according to the following composed kernel in $\mathcal{K}(\Omega_V)$:

$$K(\omega' \,|\, \omega) \;=\; \prod_{v \in V} K^{(v)}({\omega'}_v \,|\, \omega), \qquad \omega, \omega' \in \Omega_V \,.$$

(5) In applications, parallel processing is adapted to a graph $G = (V, E)$ – here, $E \subset V \times V$ denotes the set of edges – in order to model constraints for the information flow in the system. This is represented by $\mathscr{S} := \mathscr{S}(G) := \{(pa(v), \{v\}) \,:\, v \in V\}$. Each unit v is supposed to process only information from its parents $pa(v) = \{\mu \in V \,:\, (\mu, v) \in E\}$, which is modeled by a transition kernel $K^{(v)} \in \mathcal{K}(\Omega_v \,|\, \Omega_{pa(v)})$. The parallel transition of the whole system is then described by

$$K(\omega' \,|\, \omega) \;=\; \prod_{v \in V} K^{(v)}({\omega'}_v \,|\, \omega_{pa(v)}), \qquad \omega, \omega' \in \Omega_V \,.$$

(6) Now, we introduce the example of parallel processing that plays the most important role in the present paper: Consider non-empty and pairwise distinct subsystems $S_1, \ldots, S_n$ of V with $V = S_1 \uplus \cdots \uplus S_n$ and define $\mathscr{S} := \mathscr{S}(S_1, \ldots, S_n) := \{(S_i, S_i) \,:\, i = 1, \ldots, n\}$. It describes $\{S_1, \ldots, S_n\}$-*split information processing*, where the subsystems do not interact with each other. Each subsystem S_i only processes information from its own current state according to a kernel $K^{(i)} \in \mathcal{K}(\Omega_{S_i})$. The composed transition of the whole system is then given by

$$K(\omega' \,|\, \omega) \;=\; \prod_{i=1}^{n} K^{(i)}({\omega'}_{S_i} \,|\, \omega_{S_i}), \qquad \omega, \omega' \in \Omega_V \,.$$

For the completely split case, where the subsystems are the elementary units, we define $spl := \mathscr{S}(\{v\},\, v \in V) = \{(\{v\}, \{v\}) \,:\, v \in V\}$.

2.3.2. Non-Separability as Divergence from Separability

Consider a Markov chain $X_n = (X_{v,n})_{v\in V}$, $n = 0, 1, 2, \dots$, that is given by an initial distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ and a kernel $K \in \bar{\mathcal{K}}(\Omega_V)$. The probabilistic properties of this stochastic process are determined by the following set of finite marginals:

$$\begin{aligned} & \mathrm{Prob}\{X_0 = \omega_0, X_1 = \omega_1, \dots, X_n = \omega_n\} \\ = \;& p(\omega_0)\, K(\omega_1 \,|\, \omega_0) \cdots K(\omega_n \,|\, \omega_{n-1}), \qquad n = 0, 1, 2, \dots \end{aligned}$$

Thus, the set of Markov chains on Ω_V can be identified with

$$\overline{\mathrm{MC}}(\Omega_V) := \bar{\mathcal{P}}(\Omega_V) \times \bar{\mathcal{K}}(\Omega_V)$$

and we also use the notation $\{X_n\} = \{X_0, X_1, X_2, \dots\}$ instead of (p, K). The interior $\mathrm{MC}(\Omega_V)$ of the set of Markov chains carries the natural dualistic structure from $\mathcal{P}(\Omega_V \times \Omega_V)$, which is induced by the diffeomorphic composition map $\otimes : \mathrm{MC}(\Omega_V) \to \mathcal{P}(\Omega_V \times \Omega_V)$,

$$(p, K) \mapsto p \otimes K, \quad \text{with} \quad (p \otimes K)(\omega, \omega') \;:=\; p(\omega)\, K(\omega' \,|\, \omega)$$

($\otimes$ can be extended to a continuous surjective map $\overline{\mathrm{MC}}(\Omega_V) \to \bar{\mathcal{P}}(\Omega_V \times \Omega_V)$). Thus, we can talk about exponential families and (-1)-projections in $\mathrm{MC}(\Omega_V)$. The "distance" $D((p, K) \,\|\, (p', K'))$ from a Markov chain (p, K) to another one (p', K') is given by

$$D(p \otimes K \,\|\, p' \otimes K') \;=\; D(p \,\|\, p') + D_p(K \,\|\, K')\,,$$

with

$$D_p(K \,\|\, K') \;:=\; \sum_{\omega\in\Omega} p(\omega)\, D\big(K(\cdot \,|\, \omega) \,\|\, K'(\cdot \,|\, \omega)\big)\,. \tag{23}$$

For a set $\mathscr{S} = \{(A_1, B_1), (A_2, B_2), \dots, (A_n, B_n)\}$, we introduce the exponential family (see Proposition 3)

$$\mathrm{MC}_{\mathscr{S}}(\Omega_V) \;:=\; \mathcal{P}(\Omega_V) \times \mathcal{K}_{\mathscr{S}}(\Omega_V) \;\subset\; \mathrm{MC}(\Omega_V),$$

which has dimension $\big(|\Omega_V| - 1\big) + \sum_{(A,B)\in\mathscr{S}} |\Omega_A|\big(|\Omega_B| - 1\big)$.

The set of all these exponential families is partially ordered by inclusion with $\mathrm{MC}(\Omega_V)$ as the greatest element and $\mathrm{MC}_{fac}(\Omega_V)$ as the least one. This ordering is connected with the following partial ordering $\preceq$ of the sets $\mathscr{S}$: Given $\mathscr{S} = \{(A_1, B_1), \dots, (A_m, B_m)\}$ and $\mathscr{S}' = \{(A'_1, B'_1), \dots, (A'_n, B'_n)\}$, we write $\mathscr{S} \preceq \mathscr{S}'$ ($\mathscr{S}'$ *coarser* than $\mathscr{S}$) iff for all $(A, B) \in \mathscr{S}$ there exists a pair $(A', B') \in \mathscr{S}'$ with $A \subset A'$ and $B \subset B'$. One has

$$\mathscr{S} \;\preceq\; \mathscr{S}' \quad \Rightarrow \quad \mathcal{K}_{\mathscr{S}}(\Omega_V) \;\subseteq\; \mathcal{K}_{\mathscr{S}'}(\Omega_V)\,. \tag{24}$$

Thus, coarsening enlarges the corresponding manifold (the proof is given in the appendix).

Now, we describe the (-1)-projections on the exponential families $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$:

Proposition 1. *Let* (p, K) *be a Markov chain in* $\mathrm{MC}(\Omega_V)$ *and* $\mathscr{S} \preceq \mathscr{S}'$. *Then:*

(i) (PROJECTION) *The* (-1)*-projection of* (p, K) *on* $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ *is given by* $(p, K_{\mathscr{S}})$ *with* $K_{\mathscr{S}} := \otimes_{(A,B)\in\mathscr{S}} K_B^A$. *Here, the kernels* $K_B^A \in \mathcal{K}(\Omega_B \,|\, \Omega_A)$ *denote the corresponding* marginals *of* K*:*

$$K_B^A(\omega' \,|\, \omega) \;:=\; \frac{\sum_{\substack{\sigma,\sigma'\in\Omega_V \\ \sigma_A=\omega,\ \sigma'_B=\omega'}} p(\sigma)\, K(\sigma' \,|\, \sigma)}{\sum_{\substack{\sigma\in\Omega_V \\ \sigma_A=\omega}} p(\sigma)}, \qquad \omega \in \Omega_A,\ \omega' \in \Omega_B\,.$$

$K_{\mathscr{S}}$ *is the* projection of K on $\mathcal{K}_{\mathscr{S}}(\Omega_V)$ with respect to p.

(ii) (ENTROPIC REPRESENTATION) *The corresponding divergence is given by*

$$\begin{aligned} D\big((p, K) \,\|\, \mathrm{MC}_{\mathscr{S}}(\Omega_V)\big) &= D_p(K \,\|\, K_{\mathscr{S}}) \\ &= \sum_{(A,B)\in\mathscr{S}} H\big(p_A, K_B^A\big) \;-\; H(p, K)\,. \end{aligned}$$

(iii) (PYTHAGORIAN THEOREM) *One has*

$$D_p(K \,\|\, K_{\mathscr{S}}) \;=\; D_p(K \,\|\, K_{\mathscr{S}'}) + D_p(K_{\mathscr{S}'} \,\|\, K_{\mathscr{S}})\,.$$

If $K \in \mathcal{P}(\Omega_V)$, that is $K(\omega' \,|\, \omega) = p(\omega)$, $\omega, \omega' \in \Omega_V$, with a probability distribution $p \in \mathcal{P}(\Omega_V)$, then the divergence $D_p(K \,\|\, K_{fac})$ is nothing but the measure $I(p)$ for spatial interdependencies that has been discussed in the introduction and in Example 1. More generally, we interpret the divergence $D_p(K \,\|\, K_{\mathscr{S}})$ as a natural measure for the non-separability of (p, K) with respect to $\mathscr{S}$. The corresponding function $I_{\mathscr{S}} : (p, K) \mapsto I_{\mathscr{S}}(p, K) := D_p(K \,\|\, K_{\mathscr{S}})$ has a unique continuous extension to the set $\overline{\mathrm{MC}}(\Omega_V)$ of all Markov chains which is also denoted by $I_{\mathscr{S}}$ (see Lemma 4.2 in [55]). Thus, non-separability is defined for not necessarily strictly positive Markov chains.

2.4. Application to Stochastic Interaction

2.4.1. The Definition of Stochastic Interaction

As stated in the introduction we use the concept of complexity that is described by the formal definition in Equation (20) in order to define stochastic interaction.

Let V be a set of units and Ω_v, $v \in V$, corresponding state sets. Furthermore, consider non-empty and pairwise distinct subsystems $S_1, \ldots, S_n \subset V$ with $V = S_1 \uplus \cdots \uplus S_n$. The stochastic interaction of $S_1, \ldots, S_n$ with respect to $(p, K) \in \overline{\mathrm{MC}}(\Omega_V)$ is quantified by the divergence of (p, K) from the set of Markov chains that represent $\{S_1, \ldots, S_n\}$-split information processing, where the subsystems do not interact with each other (see Examples and Definitions 1 (6)). More precisely, we define the *stochastic interaction* (*of the subsystems* $S_1, \ldots, S_n$) to be the function $I_{S_1,\ldots,S_n} : \overline{\mathrm{MC}}(\Omega_V) \to \mathbb{R}_+$ with

$$I_{S_1,\ldots,S_n}(p, K) \;:=\; I_{\mathscr{S}(S_1,\ldots,S_n)}(p, K) \;=\; \inf_{K' \,\{S_1,\ldots,S_n\}\text{-split}} D_p(K \,\|\, K')\,. \tag{25}$$

In the case of complete splitting of $V = \{v_1, \ldots, v_n\}$ into the elementary units, that is $S_i := \{v_i\}$, $i = 1, \ldots, n$, we simply write I instead of $I_{\{v_1\},\ldots,\{v_n\}}$.

The definition of stochastic interaction given by Equation (25) is consistent with the complexity concept that is discussed in the introduction.

Here are some basic properties of I, which are well known in the spatial setting of Example 1:

Proposition 2. *Let V be a set of units, Ω_v, $v \in V$, corresponding state sets, and $X_n = (X_{v,n})_{v\in V}$, $n = 0, 1, 2, \ldots$, a Markov chain on Ω_V. For a subsystem $S \subset V$, we write $X_{S,n} := (X_{v,n})_{v\in S}$. Assume that the chain is given by $(p, K) \in \overline{\mathrm{MC}}(\Omega_V)$, where p is a stationary distribution with respect to K. Then the following holds:*

(i)

$$I\{X_n\} \;=\; \sum_{v\in V} H(X_{v,n+1} \,|\, X_{v,n}) \;-\; H(X_{n+1} \,|\, X_n)\,. \tag{26}$$

(ii) $A, B \subset V,\;\; A, B \neq \emptyset,\;\; A \cap B = \emptyset,\;\; A \uplus B = V \quad \Rightarrow$

$$I\{X_n\} \;=\; I\{X_{A,n}\} + I\{X_{B,n}\} + I_{A,B}\{X_n\}\,.$$

(iii) *If the process is parallel, then*

$$\begin{aligned} I\{X_n\} \;&=\; \sum_{v\in V} \Big(H(X_{v,n+1} \,|\, X_{v,n}) - H(X_{v,n+1} \,|\, X_n) \Big) \\ &=\; \sum_{v\in V} MI(X_{v,n+1}; X_{V\setminus v,n} \,|\, X_{v,n})\,. \end{aligned} \tag{27}$$

(iv) *If the process is adapted to a graph (V, E) then*

$$\begin{aligned} I\{X_n\} \;&=\; \sum_{v\in V} \Big(H(X_{v,n+1} \,|\, X_{v,n}) - H(X_{v,n+1} \,|\, X_{pa(v),n}) \Big) \\ &=\; \sum_{v\in V} MI(X_{v,n+1}; X_{pa(v)\setminus v,n} \,|\, X_{v,n})\,. \end{aligned} \tag{28}$$

In the statements (iii) and (iv), the *conditional mutual information* $MI(X; Y \,|\, Z)$ of two random variables X, Y with respect to a third one Z is defined to be the difference $H(X \,|\, Z) - H(X \,|\, Y, Z)$ (see p. 22 in [58]).

If X_{n+1} and X_n are independent for all n, the stochastic interaction $I\{X_n\}$ reduces to the measure $I(p)$ for spatial interdependencies with respect to the stationary distribution p of $\{X_n\}$ (see Example 1). Thus, the dynamical notion of stochastic interaction is a generalization of the spatial one. Geometrically, this can be illustrated as follows. In addition to the projection K_{spl} of the kernel $K \in \mathrm{MC}(\Omega_V)$ with respect to a distribution $p \in \mathcal{P}(\Omega_V)$ on the set of split kernels, we consider its projections K_{ind} and K_{fac} on the set $\mathcal{P}(\Omega_V)$ of independent kernels and on the subset $\mathcal{F}(\Omega_V)$, respectively. From Proposition 1 we know

$$\begin{aligned} D_p(K \,\|\, K_{ind}) \;&=\; H(X_{n+1}) - H(X_{n+1} \,|\, X_n)\,, \\ &\qquad (\textit{(global) transinformation}) \end{aligned}$$

$$\begin{aligned} I(p) \;=\; D_p(K_{ind} \,\|\, K_{fac}) \;&=\; \sum_{v\in V} H(X_{v,n+1}) \;-\; H(X_{n+1})\,, \\ &\qquad (\textit{spatial interdependence}) \end{aligned}$$

$$D_p(K_{spl} \,\|\, K_{fac}) \;=\; \sum_{v\in V} \big(H(X_{v,\,n+1}) - H(X_{v,\,n+1} \,|\, X_{v,\,n})\big)\,.$$
(*sum of individual transinformations*)

According to the Pythagorian relation (Proposition 1 (iii)), we get the following representation of stochastic interaction:

$$\begin{aligned} I\{X_n\} &= D_p(K \,\|\, K_{spl}) \\ &= I(p) \;+\; D_p(K \,\|\, K_{ind}) - D_p(K_{spl} \,\|\, K_{fac})\,. \end{aligned} \tag{29}$$

In the particular case of an independent process, the divergences $D_p(K \,\|\, K_{ind})$ and $D_p(K_{spl} \,\|\, K_{fac})$ in Equation (29) vanish, and the stochastic interaction coincides with spatial interdependence.

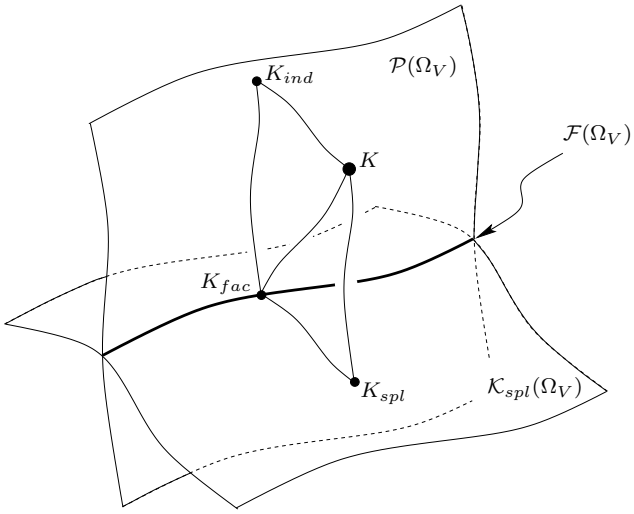

Figure 2. Illustration of the two ways of projecting K onto $\mathcal{F}(\Omega_V)$. Corresponding application of the Pythagorean theorem leads to Equation (29).

2.4.2. Examples

Example 2 (SOURCE AND RECEIVER)**.** Consider two units $1 =$ *source* and $2 =$ *receiver* with the state sets Ω_1 and Ω_2. Assume that the information flow is adapted to the graph $G = \{\{1,2\},\{(1,2)\}\}$, which only allows a transmission from the first unit to the second. In each transition from time n to $n+1$, a state $X_{1,\,n+1}$ of the first unit is chosen independently from $X_{1,\,n}$ according to a probability distribution $p \in \mathcal{P}(\Omega_1)$. The state $X_{2,\,n+1}$ of the second unit at time $n+1$ is "computed" from $X_{1,\,n}$ according to a kernel $K \in \mathcal{K}(\Omega_2 \,|\, \Omega_1)$. Using formula Equation (28), we have

$$I\{X_n\} \;=\; H(X_{2,\,n+1}) - H(X_{2,\,n+1} \,|\, X_{1,\,n})\,.$$

This is the well-known *mutual information* of the variables $X_{2,\,n+1}$ and $X_{1,\,n}$, which has a temporal interpretation within the present approach. It plays an important role in *coding and information theory* [58].

Example 3 (TWO BINARY UNITS I)**.** Consider two units with the state sets $\{0,1\}$. Each unit copies the state of the other unit with probability $1-\varepsilon$. The transition probabilities for the units are given by the following tables:

$K^{(1)}(x' \,|\, (x,y))$	0	1
$(0,0)$	$1-\varepsilon$	ε
$(0,1)$	ε	$1-\varepsilon$
$(1,0)$	$1-\varepsilon$	ε
$(1,1)$	ε	$1-\varepsilon$

$K^{(2)}(y' \,|\, (x,y))$	0	1
$(0,0)$	$1-\varepsilon$	ε
$(0,1)$	$1-\varepsilon$	ε
$(1,0)$	ε	$1-\varepsilon$
$(1,1)$	ε	$1-\varepsilon$

The transition kernel $K \in \bar{\mathcal{K}}_{par}(\{0,1\} \times \{0,1\})$ for the corresponding parallel dynamics of the whole system is then given by

$K((x',y') \,|\, (x,y))$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0)$	$(1-\varepsilon)^2$	$(1-\varepsilon)\varepsilon$	$\varepsilon(1-\varepsilon)$	ε^2
$(0,1)$	$\varepsilon(1-\varepsilon)$	ε^2	$(1-\varepsilon)^2$	$(1-\varepsilon)\varepsilon$
$(1,0)$	$(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2$	ε^2	$\varepsilon(1-\varepsilon)$
$(1,1)$	ε^2	$\varepsilon(1-\varepsilon)$	$(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2$

Note that for $\varepsilon \in \{0,1\}$, K corresponds to the deterministic transformations

$$\varepsilon = 0: \quad (x,y) \mapsto (y,x) \qquad \text{and} \qquad \varepsilon = 1: \quad (x,y) \mapsto (1-y, 1-x)\,,$$

which in an intuitive sense describe complete information exchange of the units. With the unique stationary probability distribution $p = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ one can easily compute the marginal kernels

$K_1(x' \,|\, x)$	0	1
0	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$

$K_2(y' \,|\, y)$	0	1
0	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$

which describe the split dynamics according to $K_{spl} = K_1 \otimes K_2$. With Equation (27) we finally get

$$I\{X_n\} \;=\; 2\Big(\ln 2 + (1-\varepsilon)\ln(1-\varepsilon) + \varepsilon \ln \varepsilon\Big)\,.$$

The shape of this function is shown in Figure 3.

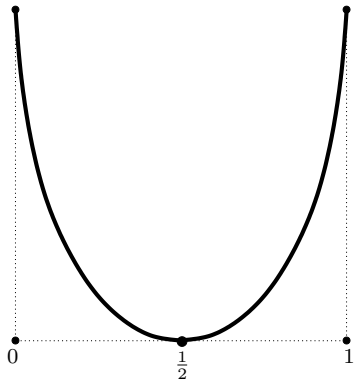

Figure 3. Illustration of the stochastic interaction $I\{X_n\}$ as a function of ε. For the extreme values of ε we have maximal stochastic interaction, which corresponds to a complete information exchange in terms of $(x, y) \mapsto (y, x)$ for $\varepsilon = 0$ and $(x, y) \mapsto (1 - y, 1 - x)$ for $\varepsilon = 1$. For $\varepsilon = \frac{1}{2}$, the dynamics is maximally random, which is associated with no interaction of the nodes.

This function is symmetric around $\varepsilon = \frac{1}{2}$ where it vanishes. In $\varepsilon = 0$ and $\varepsilon = 1$ it attains its maximal value $2 \ln 2$. As stated above, this corresponds to the deterministic transformations with complete information exchange.

Example 4 (TWO BINARY UNITS II)**.** Consider again two binary units with the state sets $\{0, 1\}$ and the transition probabilities

$K^{(1)}(x' \mid (x, y))$	0	1
$(0, 0)$	1	0
$(0, 1)$	$1 - \varepsilon$	ε
$(1, 0)$	ε	$1 - \varepsilon$
$(1, 1)$	0	1

$K^{(2)}(y' \mid (x, y))$	0	1
$(0, 0)$	0	1
$(0, 1)$	$1 - \varepsilon$	ε
$(1, 0)$	ε	$1 - \varepsilon$
$(1, 1)$	1	0

The transition kernel $K \in \bar{\mathcal{K}}(\{0, 1\} \times \{0, 1\})$ of the corresponding parallel dynamics is given by

$K((x', y') \mid (x, y))$	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
$(0, 0)$	0	1	0	0
$(0, 1)$	$(1 - \varepsilon)^2$	$(1 - \varepsilon)\varepsilon$	$\varepsilon(1 - \varepsilon)$	ε^2
$(1, 0)$	ε^2	$\varepsilon(1 - \varepsilon)$	$(1 - \varepsilon)\varepsilon$	$(1 - \varepsilon)^2$
$(1, 1)$	0	0	1	0

Note that for $\varepsilon \in \{0, 1\}$, K corresponds to the deterministic transformations

$$\varepsilon = 0 : \quad (x, y) \mapsto (x, 1 - y) \qquad \text{and} \qquad \varepsilon = 1 : \quad (x, y) \mapsto (y, 1 - x) \, .$$

Thus in an intuitive sense, for $\varepsilon = 1$ the units completely interact with each other, and for $\varepsilon = 0$ there is no interaction. For $\varepsilon \in]0,1[$ we compute the interaction with respect to the unique stationary probability distribution

$$p = \tfrac{1}{4(\varepsilon^2-\varepsilon+1)}\,(2\varepsilon^2-2\varepsilon+1\ ,\ 1\ ,\ 1\ ,\ 2\varepsilon^2-2\varepsilon+1)\,.$$

With the corresponding marginal kernels

$K_1(x' \mid x)$	0	1
0	$1-\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$
1	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$1-\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$

$K_2(y' \mid y)$	0	1
0	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$1-\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$
1	$1-\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$

and Equation (27), we get

$$\begin{aligned} I\{X_n\} \;=\; \frac{\varepsilon}{\varepsilon^2-\varepsilon+1}\Big(&-(2\varepsilon^2-3\varepsilon+2)\ln(2\varepsilon^2-3\varepsilon+2) \\ &+2(\varepsilon^2-\varepsilon+1)\ln 2(\varepsilon^2-\varepsilon+1) \;+\; (1-\varepsilon)\ln(1-\varepsilon)\Big)\,. \end{aligned}$$

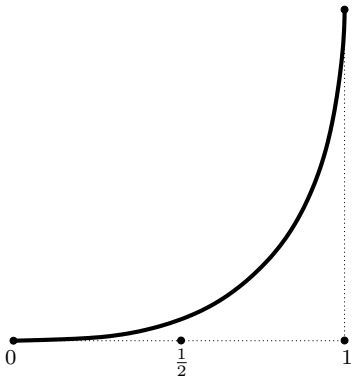

Figure 4. Illustration of the stochastic interaction $I\{X_n\}$ as a function of ε. For $\varepsilon = 0$, the two units update their states with no information exchange: $(x,y) \mapsto (x, 1-y)$. For $\varepsilon = 1$, there is maximal information exchange in terms of $(x,y) \mapsto (y, 1-x)$.

This function is monotonically increasing from the minimal value 0 (no interaction) in $\varepsilon = 0$ to its maximal value $2\ln 2$ (complete interaction) in $\varepsilon = 1$.

3. Conclusions

Following the general concept that complexity is characterized by the divergence of a composed system from the superposition of its elementary parts, information geometry has been used to derive a measure for spatio-temporal interdependencies among a finite set of units, which is referred to as *stochastic interaction*. This generalizes the well-known measure for spatial interdependence that is quantified by the Kullback-Leibler divergence of a probability distribution from its factorization [18,55]. Thereby, previous work by Ay [23] is continued, where the optimization of dependencies among

stochastic units has been proposed as a principle for neural organization in feed-forward networks. Of course, the present setting is much more general and provides a way to consider also recurrent networks. The dynamical properties of strongly interacting units in the sense of the present paper are studied by Ay and Wennekers in [24], where the emergence of determinism and structure in such systems is demonstrated.

Appendix: Proofs

Proposition 3. *The manifold* $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ *is an exponential family in* $\mathrm{MC}(\Omega_V)$.

Proof. To see this, consider the functions $\Omega_V \times \Omega_V \to \mathbb{R}$

$$v_\sigma(\omega,\omega') := \begin{cases} 1, & \text{if } \omega = \sigma \\ 0, & \text{otherwise} \end{cases}, \qquad \sigma \in \Omega_V\,,$$

and

$$v_{\sigma,\sigma'}(\omega,\omega') := \begin{cases} 1, & \text{if } \omega_A = \sigma,\ \omega'_B = \sigma' \\ 0, & \text{otherwise} \end{cases}, \qquad (A,B) \in \mathscr{S},\ \sigma \in \Omega_A,\ \sigma' \in \Omega_B\,.$$

It is easy to verify that the image of $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ under the map $\otimes$ is the following exponential family in $\mathcal{P}(\Omega_V \times \Omega_V)$:

$$\exp\left\{\sum_{\sigma\in\Omega_V} \lambda_\sigma\, v_\sigma + \sum_{(A,B)\in\mathscr{S}} \sum_{\sigma\in\Omega_A,\,\sigma'\in\Omega_B} \lambda_{\sigma,\sigma'}\, v_{\sigma,\sigma'} \;-\; \Theta\right\}, \qquad \lambda_\sigma, \lambda_{\sigma,\sigma'} \in \mathbb{R}\,.$$

Here, Θ denotes the normalization factor, which depends on the λ-parameters. In particular, each element in $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ can be expressed in the following way

$$\begin{aligned} & p(\omega) \prod_{(A,B)\in\mathscr{S}} K^A_B(\omega'_B \,|\, \omega_A) \\ = \;& \exp\left\{\ln p(\omega) + \sum_{(A,B)\in\mathscr{S}} \ln K^A_B(\omega'_B \,|\, \omega_A)\right\} \\ = \;& \exp\left\{\sum_{\sigma\in\Omega_V} \ln p(\sigma)\; v_\sigma(\omega,\omega') + \sum_{(A,B)\in\mathscr{S}} \sum_{\sigma\in\Omega_A,\,\sigma'\in\Omega_B} \ln K^A_B(\sigma' \,|\, \sigma)\; v_{\sigma,\sigma'}(\omega,\omega')\right\}. \end{aligned}$$

□

Proof of Implication (24). If

$$\mathscr{S} = \{(A_1,B_1),\ldots,(A_m,B_m)\} \;\preceq\; \mathscr{S}'\{(A'_1,B'_1),\ldots,(A'_n,B'_n)\}\,,$$

then there exists a partition M_i, $i = 1,\ldots,n$, of the index set $\{1,\ldots,m\}$ such that

$$B'_i \;=\; \biguplus_{j\in M_i} B_j, \qquad i = 1,\ldots,n\,.$$

Let (p, K) be a Markov chain in $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$. Then there exist $K_B^A \in \mathcal{K}(\Omega_B \,|\, \Omega_A)$ with

$$\begin{aligned} K(\omega' \,|\, \omega) &= \prod_{(A,B)\in\mathscr{S}} K_B^A(\omega'_B \,|\, \omega_A) \\ &= \prod_{i=1}^{n} \underbrace{\prod_{\substack{(A_j,B_j)\in\mathscr{S} \\ j\in M_i}} K_{B_j}^{A_j}(\omega'_{B_j} \,|\, \omega_{A_j})}_{=:\, K_{B'_i}^{A'_i}(\omega'_{B'_i} \,|\, \omega_{A'_i})}, \qquad \omega, \omega' \in \Omega_V \,. \end{aligned}$$

The kernels $K_{B'_i}^{A'_i}$ are contained in $\mathcal{K}_{\mathscr{S}'}$, and therefore we get $(p, K) \in \mathrm{MC}_{\mathscr{S}'}(\Omega_V)$. □

Proof of Proposition 1.

(i) Consider the following strictly convex function ($\mathbb{R}_+^*$ denotes the set of positive real numbers)

$$F: \left(\mathbb{R}_+^*\right)^{\Omega_V} \times \left(\prod_{(A,B)\in\mathscr{S}} \left(\mathbb{R}_+^*\right)^{\Omega_A\times\Omega_B} \right) \to \mathbb{R} \,,$$

$$(x, y) = (x_\omega,\, \omega \in \Omega_V; y_{\omega_A,\omega_B},\, \omega_A \in \Omega_A, \omega_B \in \Omega_B) \mapsto$$

$$\begin{aligned} F(x,y) \;:=\; & \sum_{\omega\in\Omega_V} p(\omega) \ln \frac{p(\omega)}{x_\omega} \;+\; \sum_{\omega,\,\omega'\in\Omega_V} p(\omega)\, K(\omega' \,|\, \omega) \ln \frac{K(\omega' \,|\, \omega)}{\prod_{(A,B)\in\mathscr{S}} y_{\omega_A,\omega'_B}} \\ & + \lambda \left(\sum_{\omega\in\Omega_V} x_\omega \;-\; 1 \right) + \sum_{(A,B)\in\mathscr{S}} \sum_{\omega_A\in\Omega_A} \lambda_{\omega_A}^B \left(\sum_{\omega'_B\in\Omega_B} y_{\omega_A,\omega'_B} - 1 \right) . \end{aligned}$$

Here, λ and the $\lambda_{\omega_A}^B$ are Lagrangian parameters. Note that in the case $x \in \mathcal{P}(\Omega_V)$ and $y \in \prod_{(A,B)\in\mathscr{S}} \mathcal{K}(\Omega_B \,|\, \Omega_A)$, the value $F(x, y)$ is nothing but the divergence of (p, K) from $(x, \otimes_{\mathscr{S}}(y))$. In order to get the Markov chain that minimizes the divergence we have to compute the partial derivatives of F:

$$\begin{aligned} \frac{\partial F}{\partial x_\sigma}(x,y) &= -\sum_{\omega\in\Omega_V} p(\omega)\, \frac{1}{x_\omega}\, \delta_{\sigma,\,\omega} \;+\; \lambda \\ &= -\frac{p(\sigma)}{x_\sigma} \;+\; \lambda \,, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial F}{\partial y_{\sigma_C,\sigma'_D}}(x,y) &= -\sum_{\omega,\omega'\in\Omega_V} p(\omega)\, K(\omega' \,|\, \omega) \sum_{(A,B)\in\mathscr{S}} \frac{1}{y_{\omega_A,\omega'_B}} \delta_{(\omega_A,\omega'_B),(\sigma_C,\sigma'_D)} \\ &\quad + \sum_{(A,B)\in\mathscr{S}} \sum_{\omega_A\in\Omega_A} \lambda_{\omega_A}^B \sum_{\omega'_B\in\Omega_B} \delta_{(\omega_A,\omega'_B),(\sigma_C,\sigma'_D)} \\ &= -\sum_{\omega,\omega'\in\Omega_V} p(\omega)\, K(\omega' \,|\, \omega) \frac{1}{y_{\omega_C,\omega'_D}} \delta_{(\omega_C,\omega'_D),(\sigma_C,\sigma'_D)} \;+\; \lambda_{\sigma_C}^D \\ &= -\frac{1}{y_{\sigma_C,\sigma'_D}} \sum_{\substack{\omega,\omega'\in\Omega_V \\ \omega_C=\sigma_C,\, \omega'_D=\sigma'_D}} p(\omega)\, K(\omega' \,|\, \omega) \;+\; \lambda_{\sigma_C}^D \,. \end{aligned}$$

For a critical point (x, y), the partial derivatives vanish. We get the following solution:

$$x_\sigma \;=\; p(\sigma), \qquad \sigma \in \Omega_V\,,$$

and

$$y_{\sigma_C,\sigma'_D} \;=\; \frac{1}{\sum_{\omega_C\in\Omega_C} p(\omega)} \sum_{\substack{\omega,\omega'\in\Omega_V \\ \omega_C=\sigma_C,\,\omega'_D=\sigma'_D}} p(\omega)\,K(\omega'\,|\,\omega) \qquad \sigma_C\in\Omega_C,\;\sigma'_D\in\Omega_D\,.$$

From Theorem 3.10 in [17] we know that this solution is the (-1)-projection of (p, K) onto $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$. It is given by the initial distribution p and the corresponding marginals K^A_B, $(A, B) \in \mathscr{S}$, of K.

(ii) With (i) we get

$$\begin{aligned}
& D\big((p,K)\,\|\,\mathrm{MC}_{\mathscr{S}}(\Omega_V)\big) \\
=\;& D_p(K\,\|\,K_{\mathscr{S}}) \\
=\;& \sum_{\omega,\omega'\in\Omega_V} p(\omega)\,K(\omega'\,|\,\omega)\ln\frac{K(\omega'\,|\,\omega)}{\prod_{(A,B)\in\mathscr{S}} K^A_B(\omega'_B\,|\,\omega_A)} \\
=\;& -H(p,K) \\
& -\sum_{(A,B)\in\mathscr{S}}\sum_{\omega,\omega'\in\Omega_V} p(\omega)\,K(\omega'\,|\,\omega)\ln K^A_B(\omega'_B\,|\,\omega_A) \\
=\;& -H(p,K) \\
& -\sum_{(A,B)\in\mathscr{S}}\sum_{\omega\in\Omega_A,\omega'\in\Omega_B} \ln K^A_B(\omega'\,|\,\omega)\underbrace{\sum_{\substack{\sigma,\sigma'\in\Omega_V \\ \sigma_A=\omega,\;\sigma'_B=\omega'}} p(\sigma)\,K(\sigma'\,|\,\sigma)}_{p_A(\omega)\,K^A_B(\omega'\,|\,\omega)} \\
=\;& \sum_{(A,B)\in\mathscr{S}} H(p_A,K^A_B)\;-\;H(p,K).
\end{aligned}$$

(iii) According to Equation (24) we have $\mathrm{MC}_{\mathscr{S}}(\Omega_V)\subseteq\mathrm{MC}_{\mathscr{S}'}(\Omega_V)$, and the statement follows from the Pythagorian theorem ([17], p. 62, Theorem 3.8). □

Proof of Proposition 2.

(i) This follows from Proposition 1 (ii).

(ii) We apply (i):

$$\begin{aligned}
I\{X_n\} \;&\overset{\text{(i)}}{=}\; \sum_{v\in V} H(X_{v,n+1}\,|\,X_{v,n}) \;-\; H(X_{n+1}\,|\,X_n) \\
&= \left(\sum_{v\in A} H(X_{v,n+1}\,|\,X_{v,n}) \;-\; H(X_{A,n+1}\,|\,X_{A,n})\right) \\
&\quad + \left(\sum_{v\in B} H(X_{v,n+1}\,|\,X_{v,n}) \;-\; H(X_{B,n+1}\,|\,X_{B,n})\right) \\
&\quad + \Big(H(X_{A,n+1}\,|\,X_{A,n}) + H(X_{B,n+1}\,|\,X_{B,n}) - H(X_{n+1}\,|\,X_n)\Big) \\
&\overset{\text{(i)}}{=}\; I\{X_{A,n}\} + I\{X_{B,n}\} + I_{A,B}\{X_n\}\,.
\end{aligned}$$

(iii) For parallel processing, one has

$$H(X_{n+1} \,|\, X_n) \;=\; \sum_{v \in V} H(X_{v,\,n+1} \,|\, X_n)\,.$$

The statement is then implied by (i).

(iv) This follows from (iii) and the Markov property for (V, E)-adapted Markov chains. □

Conflicts of Interest

The author declares no conflict of interest.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Attneave, F. Informational aspects of visual perception. *Psychol. Rev.* **1954**, *61*, 183–193.
3. Barlow, H.B. Possible principles underlying the transformation of sensory messages. *Sens. Commun.* **1961**, 217–234.
4. Laughlin, S. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch.* **1981**, *36*, 910–912.
5. Linsker, R. Self-organization in a perceptual network. *Computer* **1988**, *21*, 105–117.
6. Hubel, D.H.; Wiesel, T.N. Functional Architecture of Macaque Monkey Visual Cortex (Ferrier lecture). *Proc. R. Soc. Lond. B* **1977**, *198*, 1–59.
7. Hubel, D.H.; Wiesel, T.N. Brain Mechanisms of Vision. *Sci. Am.* **1979**, *241*, 150–162.
8. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **1995**, *7*, 1129–1159.
9. Tononi, G.; Sporns, O.; Edelmann, 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.
10. Tononi, G.; Sporns, O. Measuring information integration. *BMC Neurosci.* **2003**, *4*, doi:10.1186/1471-2202-4-31.
11. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **2004**, *5*, doi:10.1186/1471-2202-5-42.
12. Balduzzi, D.; Tononi, G. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Comp. Biol.* **2008**, *4*, 1000091.
13. Tononi, G. Consciousness as Integrated Information: A Provisional Manifesto. *Biol. Bull.* **2008**, *215*, 216–242.
14. Barrett, A.B.; Seth, A.K. Practical Measures of Integrated Information for Time-Series Data. *PLoS Comp. Biol.* **2011**, *7*, 1001052.
15. Edlund, J.; Chaumont, N.; Hintze, A.; Koch, C.; Tononi, G.; Adami, C. Integrated Information Increases with Fitness in the Evolution of Animats. *PLoS Comp. Biol.* **2011**, *7*, e1002236.
16. Ay, N. Information Geometry on Complexity and Stochastic Interaction. MPI MiS Preprint, No. 95/2001. Available online: http://www.mis.mpg.de/publications/preprints/2001/prepr2001-95.html (accessed on 7 December 2001).

17. Amari, S.I.; Nagaoka, H. *Methods of Information Geometry*; Translations of Mathematical Monographs; AMS and Oxford University Press: New York, NY, USA, 2000.
18. Amari, S.I. Information Geometry on Hierarchy of Probability Distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711.
19. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
20. Kolchinsky, A.; Rocha, L.M. Prediction and Modularity in Dynamical Systems. In *Advances in Artificial Life, Proceedings of the Eleventh European Conference on the Syntheses and Simulation of Living Systems (ECAL 2011)*; MIT Press: Paris, France, 2011; pp. 423–430.
21. Arsiwalla, X.D.; Verschure, P.F.M.J. Integrated Information for Large Complex Networks. In Proceedings of 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 620–626.
22. Amari, S.I. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276.
23. Ay, N. Locality of global stochastic interaction in directed acyclic networks. *Neural Comput.* **2002**, *14*, 2959–2980.
24. Ay, N.; Wennekers, T. Dynamical Properties of Strongly Interacting Markov Chains. *Neural Netw.* **2003**, *16*, 1483–1497.
25. Wennekers, T.; Ay, N. Finite state automata resulting from temporal information maximization and a temporal learning rule. *Neural Comput.* **2005**, *17*, 2258–2290.
26. Ay, N.; Montúfar, G.; Rauh, J. Selection Criteria for Neuromanifolds of Stochastic Dynamics. In *Advances in Cognitive Neurodynamics (III)*, Proceedings of the Third International Conference on Cognitive Neurodynamics 2011, Hokkaido, Japan, 9–13 June 2011; pp. 147–154.
27. Ay, N. *Geometric Design Principles for Brains of Embodied Agents*; Santa Fe Institute Working Paper 15-02-005; Santa Fe Institute: Santa Fe, NM, USA, 2015.
28. Wennekers, T.; Ay, N. Temporal infomax leads to almost deterministic dynamical systems. *Neurocomputing* **2003**, *52-4*, 461–466.
29. Wennekers, T.; Ay, N. Temporal Infomax on Markov chains with input leads to finite state automata. *Neurocomputing* **2003**, *52-4*, 431–436.
30. Wennekers, T.; Ay, N. Spatial and temporal stochastic interaction in neuronal assemblies. *Theory Biosci.* **2003**, *122*, 5–18.
31. Wennekers, T.; Ay, N. Stochastic interaction in associative nets. *Neurocomputing* **2005**, *65*, 387–392.
32. Wennekers, T.; Ay, N. A temporal learning rule in recurrent systems supports high spatio-temporal stochastic interactions. *Neurocomputing* **2006**, *69*, 1199–1202.
33. Wennekers, T.; Ay, N.; Andras, P. High-resolution multiple-unit EEG in cat auditory cortex reveals large spatio-temporal stochastic interactions. *Biosystems* **2007**, *89*, 190–197.
34. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.
35. Ay, N.; Polani, D. Information Flows in Causal Networks. *Adv. Complex Syst.* **2008**, *11*, 17–41.
36. Ay, N.; Krakauer, D.C. Geometric robustness theory and biological networks. *Theory Biosci.* **2007**, *2*, 93–121.
37. Ay, N. A Refinement of the Common Cause Principle. *Discret. Appl. Math.* **2009**, *157*, 2439–2457.

38. Steudel, B.; Ay, N. Information-theoretic inference of common ancestors. *Entropy* **2015**, *17*, 2304–2327.
39. Moritz, P.; Reichardt, J.; Ay, N. Discriminating between causal structures in Bayesian Networks via partial observations. *Kybernetika* **2014**, *50*, 284–295.
40. Pearl, J. *Causality: Models, Reasoning and Inference*; Cambridge University Press: Cambridge, UK, 2000.
41. Janzing, D.; Balduzzi, D.; Grosse-Wentrup, M.; Schölkopf, B. Quantifying causal influences. *Ann. Stat.* **2013**, *41*, 2324–2358.
42. Ay, N.; Knauf, A. Maximizing Multi-Information. *Kybernetika* **2007**, *42*, 517–538.
43. Ay, N.; Olbrich, E.; Bertschinger, N.; Jost, J. A Geometric Approach to Complexity. *Chaos* **2011**, *21*, 037103.
44. Ay, N.; Olbrich, E.; Bertschinger, N.; Jost, J. A Unifying Framework for Complexity Measures of Finite Systems. In Proceedings of the European Conference on Complex Systems 2006 (ECCS'06), Oxford University, Oxford, UK, 25–29 September 2006; p. 80.
45. Adami, C. The Use of Information Theory in Evolutionary Biology. *Ann. NY Acad. Sci.* **2012**, *1256*, 49–65.
46. Ay, N.; Bertschinger, N.; Der, R.; Guettler, F.; Olbrich, E. Predictive information and explorative behavior of autonomous robots. *Eur. Phys. J. B* **2008**, *63*, 329–339.
47. Zahedi, K.; Ay, N.; Der, R. Higher coordination with less control—A result of information maximisation in the sensorimotor loop. *Adapt. Behav.* **2010**, *18*, 338–355.
48. Ay, N.; Bernigau, H.; Der, R.; Prokopenko, M. Information driven self-organization: The dynamical system approach to autonomous robot behavior. *Theory Biosci.* **2011**. doi:10.1007/s12064-011-0137-9.
49. Kahle, T.; Olbrich, E.; Jost, J.; Ay, N. Complexity Measures from Interaction Structures. *Phys. Rev. E* **2009**, *79*, 026201.
50. Olbrich, E.; Bertschinger, N.; Ay, N.; Jost, J. How should complexity scale with system size? *Eur. Phys. J. B* **2008**, *63*, 407–415.
51. Amari, S.I. *Differential-Geometric Methods in Statistics (Lecture Notes in Statistics)*; Springer: Berlin, Germany, 1985.
52. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
53. Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158.
54. Csiszár, I. On topological properties of f-divergence. *Stud. Sci. Math. Hungar* **1967**, *2*, 329–339.
55. Ay, N. An Information-Geometric Approach to a Theory of Pragmatic Structuring. *Ann. Probab.* **2002**, *30*, 416–436.
56. Rao, C.R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
57. Hofbauer, J.; Sigmund, K. *Evolutionary Games and Population Dynamics*; Cambridge University Press: Cambridge, UK, 1998.

58. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley Series in Telecommunications; Wiley-Interscience: New York, NY, USA, 1991.