Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

A Geometric Approach to Complexity

by

Nihat Ay, Eckehard Olbrich, Nils Bertschinger, and Jürgen Jost

Preprint no.: 53 2011

A Geometric Approach to Complexity

Nihat Ay[1,2], Eckehard Olbrich[1], Nils Bertschinger[1], Jürgen Jost[1,2]

{nay, bertschi, jost, olbrich}@mis.mpg.de

[1]Max Planck Institute for Mathematics in the Sciences
Inselstrasse 22, 04103 Leipzig, Germany

[2]Santa Fe Institute
1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

Abstract: We develop a geometric approach to complexity based on the principle that complexity requires interactions at different scales of description. Complex systems are more than the sum of their parts of any size, and not just more than the sum of their elements. Using information geometry, we therefore analyze the decomposition of a system in terms of an interaction hierarchy. In mathematical terms, we present a theory of complexity measures for finite random fields using the geometric framework of hierarchies of exponential families. Within our framework, previously proposed complexity measures find their natural place and gain a new interpretation.

Keywords: complexity, information geometry, hierarchical model, predictive information, excess entropy

Lead paragraph: Various complexity measures for composite systems have been proposed. Each of them has its own conceptual and theoretical justification. It is desirable to have a framework for the systematic comparison of such complexity measures that can provide a unified view with corresponding analytical results. Some results within this line of research are known, for examples in the context of statistical complexity and excess entropy. We develop a geometric approach to complexity which allows us to extend this line of research. We believe that this approach is very natural and rich, and we demonstrate its utility in this regard by deriving analytical results related to the complexity measure proposed by Tononi, Sporns, and Edelman, and also to excess entropy. Both complexity measures are well-known and turned out to be natural and useful. Therefore, the possibility of discussing them from a unified perspective in terms of our formalism appears very promising and is subject of our ongoing research.

1. Introduction

The most famous quote about complex systems is attributed to Aristotle and says 'The whole is more than the sum of its parts'.[1] Complex systems are systems where the collective behavior of their parts entails emergence of properties that can hardly, if not at all, be inferred from properties of the parts. In this article we draw a geometric picture that allows us to approach a formalization of this concept. Building on previous work [2, 8, 4, 28], we propose information geometry [1] as one instance of this approach.

[1]In fact, this is an abbreviation of a longer reasoning of Aristotle in his *Metaphysics*, Vol.VII, 1041b.

There is quite a number of formal approaches to complexity, and currently there is no unifying theory that incorporates all of them. On the other hand, at an intuitive level, they arise from a few ideas and most approaches are based on at least one of the following three conceptual lines, which assume complexity to be the[2] ...

(1) ... minimal effort that one has to make, or minimal resources that one has to use, in order to describe or generate an object. Examples are: algorithmic complexity (Kolmogorov [25]; Chaitin [11]; Solomonoff [36]), computational complexities [21, 31], entropy or entropy rate (Shannon [33]; Kolmogorov [26, 27]; Sinai [34]).
(2) ... minimal effort that one has to make, or minimal resources that one has to use, in order to describe or generate *regularities* or *structure* of an object. Examples are: Kolmogorov minimal sufficient statistics and related notions (see [27]), stochastic complexity (Rissanen [32]), effective complexity (Gell-Mann and Lloyd [17], [18], [5], [6]), statistical complexity (Crutchfield and Young [13]; Shalizi and Crutchfield [35]), excess entropy (Grassberger [19]; Shalizi and Crutchfield [35]; Bialek, Nemenman, and Tishby [10]).
(3) ... the extent to which an object, as a whole, is more than the sum of its parts. The extent to which the whole cannot be understood by analysis of the system parts alone, but only by taking into account their interactions.

The third conceptual line is commonly assumed to be fundamental. However, only a few attempts have been made towards a formalization of this property. One approach that goes in that direction is due to Tononi, Sporns, and Edelman [37] and provides a complexity measure which we refer to as *TSE-complexity*. A more explicit attempt to formalize the third concept has been made in [3] using information geometry. In this article we extend this line of research and relate it to other known approaches. The information-geometric framework provides us with a new view on the TSE-complexity and allows us to relate this complexity measure to a larger class (Section 3.1). At the same time the information-geometric perspective yields a new interpretation of the excess entropy (Section 3.2).

2. A Geometric Approach to Complexity

2.1. **The General Idea.** In this article we develop a geometric formalization of the idea that in a complex system the whole is more than the sum of its parts. We start with a few very general arguments sketching the geometric picture. However, in this sketch the formal objects are not introduced explicitly. We will specify this approach within the framework of information geometry in Section 2.2.

Assume that we have given a set $\mathcal{S}$ of systems to which we want to assign a complexity measure. This set can in principle refer to any of the above-mentioned levels of description. In order to study the complexity of a system $s \in \mathcal{S}$ we have to compare it with the sum of its parts. This requires that we have a notion of system parts. The specification of the parts of a given system is a fundamental problem in systems theory. In various approaches one considers all partitions (or coverings) that are optimal in a sense that is appropriate within the given context. We do not explicitly address this problem here but consider several choices of system parts. Each collection of system parts that is assigned to a given system s may be an element of a set $\mathcal{D}$ that formally differs from $\mathcal{S}$. We interpret the corresponding assignment $\mathsf{D} : \mathcal{S} \to \mathcal{D}$ as a reduced description of the system in terms of its parts. Having the parts $\mathsf{D}(s)$ of a system s, we have to reconstruct s,

[2]Note that the reference to particular theories is by no means complete.

that is take the "sum of the parts", in order to obtain a system that can be compared with the original system. The corresponding construction map is denoted by $\mathsf{C} : \mathcal{D} \to \mathcal{S}$. The composition

$$\mathsf{P}(s) := (\mathsf{C} \circ \mathsf{D})(s)$$

then corresponds to the sum of parts of the system s, and we can compare s with $\mathsf{P}(s)$. If these two systems coincide, then we would say that the whole equals the sum of its parts and therefore does not display any complexity. We refer to these systems as non-complex systems and denote their set by $\mathcal{N}$, that is $\mathcal{N} = \{s \in \mathcal{S} \ : \ \mathsf{P}(s) = s\}$. Note that the utility of this idea within complexity theory strongly depends on the concrete specification of the maps D and C. As mentioned above, the right definition of D incorporates the identification of system parts which is already a fundamental problem. Furthermore, one can think of specifications of D and C that do not reflect common intuitions of complexity.

The above representation of non-complex systems is implicit, and we now derive an explicit one. Obviously, we have

$$\mathcal{N} \ \subseteq \ \mathsf{im}(\mathsf{P}) \ = \ \{\mathsf{P}(s) \ : \ s \in \mathcal{S}\}.$$

With the following natural assumption, we even have equality, which provides an explicit represnetation of non-complex systems as the image of the map P. We assume that the construction of a system from a set of parts in terms of C does not generate any additional structure. More formally,

$$(\mathsf{D} \circ \mathsf{C})(s') \ = \ s' \qquad \text{for all } s' \in \mathcal{D}.$$

This implies

$$\begin{aligned} \mathsf{P}^2 &= (\mathsf{C} \circ \mathsf{D}) \circ (\mathsf{C} \circ \mathsf{D}) \\ &= \mathsf{C} \circ (\mathsf{D} \circ \mathsf{C}) \circ \mathsf{D} \\ &= \mathsf{C} \circ \mathsf{D} \\ &= \mathsf{P}. \end{aligned}$$

Thus, the above assumption implies that P is idempotent and we can interpret it as a projection. This yields the explicit representation of the set $\mathcal{N}$ of non-complex systems as the image of P.

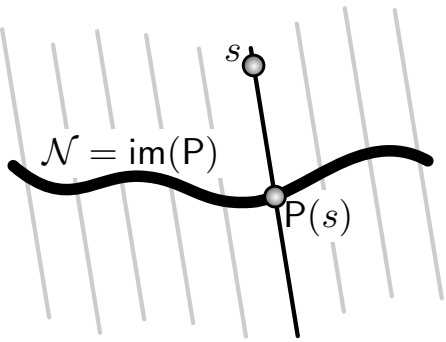

FIGURE 1. Illustration of the projection P that assigns to each system s the sum of its parts, the system $\mathsf{P}(s)$. The image of this projection constitutes the set of non-complex systems.

In order to have a quantitative theory of complexity one needs a deviation measure $d : \mathcal{S} \times \mathcal{S} \to \overline{\mathbb{R}}$ which allows us to measure the deviation of the system s from $\mathsf{P}(s)$, the sum of its parts. We

assume $d(s,s') \geq 0$, and $d(s,s') = 0$ if and only if $s = s'$. In order to ensure compatibility with P, one has to further assume that d satisfies

$$C(s) \;:=\; d(s,\mathsf{P}(s)) \;=\; \inf_{s' \in \mathcal{N}} d(s,s'). \tag{1}$$

Obviously, the complexity $C(s)$ of a system s vanishes if and only if the system s is an element of the set $\mathcal{N}$ of non-complex systems.

2.2. **The Information-Geometric Approach.**

2.2.1. *A specification via the maximum entropy method.* In this section we want to link the general geometric idea of Section 2.1 to a particular formal setting. In order to have a notion of parts of a system, we assume that the system consists of a finite node set V and that each node v can be in finitely many states $\mathscr{X}_v$. We model the whole system by a probability measure p on the corresponding product configuration set $\mathscr{X}_V = \prod_{v \in V} \mathscr{X}_v$. The parts are given by marginals p_A where A is taken from a set $\mathfrak{A}$ of subsets of V. For convenience, we will assume that $\mathfrak{A}$ satisfies the conditions of the following definition.

Definition 2.1. We call a subset $\mathfrak{A}$ of the power set 2^V a *simplicial complex* if it satisfies

(1) $\bigcup_{A \in \mathfrak{A}} A = V$, and
(2) $A \in \mathfrak{A},\ B \subseteq A \quad \Rightarrow \quad B \in \mathfrak{A}$.

♦

The decomposition of a probability measure p into parts, corresponding to D in Section 2.1, is given by the map $p \to (p_A)_{A \in \mathfrak{A}}$, which assigns to p the family of the marginals. The reconstruction of the system as a whole from this information is naturally given by the maximum entropy estimate of p, which serves as the map C in Section 2.1. To be more precise, we consider the Shannon entropy, assigned to a random variable X which assumes values x in a finite set $\mathscr{X}$. Denoting the distribution of X by p, it is defined as

$$H_p(X) \;:=\; -\sum_{x \in \mathscr{X}} p(x) \log_2 p(x). \tag{2}$$

Given a family p_A, $A \in \mathfrak{A}$, of marginals, by $\mathcal{M}$ we denote the set of probability measures q on $\mathscr{X}_V$ that have the same marginals, that is $q_A = p_A$, $A \in \mathfrak{A}$. From the strict concavity of the Shannon entropy (2) as a function of p, it follows that there is a unique probability measure $p^* \in \mathcal{M}$ with maximal entropy. This corresponds to $\mathsf{P}(p)$, and it is obvious that in this particular case $\mathsf{P}^2 = \mathsf{P}$. If the original system coincides with its maximum entropy estimate $p^* = \mathsf{P}(p)$ then p is nothing but the "sum of its parts" and is interpreted as a non-complex system, that is, p is an element of the set $\mathcal{N}$ in Section 2.1. What does this set look like in the context of maximum entropy estimation? This is simply the image of the map P, that is, all the maximum entropy distributions with given marginals on the $A \in \mathfrak{A}$, which is known as the closure of a *hierarchical model*. In the next Section 2.2.2, we will describe this set of non-complex systems in more detail.

As mentioned in Section 2.1, in order to quantify complexity, we need a deviation measure that is compatible with the projection P, here the maximum entropy estimation. Information geometry suggests to use the *relative entropy*, also called *KL-divergence*. For two given probability measures

p and q, it is defined as follows:

$$D(p \,\|\, q) \;:=\; \begin{cases} \sum_{x \in \mathscr{X}} p(x) \log_2 \frac{p(x)}{q(x)}, & \text{if } \mathsf{supp}(p) \subseteq \mathsf{supp}(q) \\ \infty \qquad\qquad\quad , & \text{otherwise.} \end{cases}$$

The following version of the consistency property (1) follows from general results of information geometry [7]:

$$D(p \,\|\, \mathsf{P}(p)) \;=\; \inf_{q \,\in\, \mathsf{im}(\mathsf{P})} D(p \,\|\, q).$$

2.2.2. *Interaction Spaces and Hierarchical Models.* In this section we describe the image of the maximum entropy projection. According to our interpretation, this consists of those systems that do not display any structure in addition to the one given by the parts.

For every subset $A \subseteq V$, the configurations on A are given by the Cartesian product

$$\mathscr{X}_A \;:=\; \prod_{v \in A} \mathscr{X}_v.$$

Note that in the case where A is the empty set, the product space consists of the empty sequence ε, that is $\mathscr{X}_\emptyset = \{\varepsilon\}$. We have the natural projections

$$X_A : \mathscr{X}_V \to \mathscr{X}_A, \qquad (x_v)_{v \in V} \mapsto (x_v)_{v \in A}.$$

Given a distribution p, the X_A become random variables and we denote the X_A-image of p by p_A, which is the A-marginal of p.

We decompose $x \in \mathscr{X}_V$ as $x = (x_A, x_{V \setminus A})$ with $x_A \in \mathscr{X}_A$, $x_{V \setminus A} \in \mathscr{X}_{V \setminus A}$, and define $\mathcal{I}_A$ to be the subspace of functions that do not depend on the configurations $x_{V \setminus A}$:

$$\begin{aligned} \mathcal{I}_A \;:=\; \Big\{ f \in \mathbb{R}^{\mathscr{X}_V} \;:\; & f(x_A, x_{V \setminus A}) = f(x_A, x'_{V \setminus A}) \\ & \text{for all } x_A \in \mathscr{X}_A, \text{ and all } x_{V \setminus A},\, x'_{V \setminus A} \in \mathscr{X}_{V \setminus A} \Big\}. \end{aligned}$$

This is called the space of *A-interactions*. Clearly, this space has dimension $\prod_{v \in A} |\mathscr{X}_v|$. The vector space of *pure A-interactions* is defined as

$$\widetilde{\mathcal{I}}_A \;:=\; \mathcal{I}_A \cap \left(\bigcap_{B \subsetneq A} \mathcal{I}_B{}^{\perp} \right).$$

Here, the orthogonal complements are taken with respect to the canonical scalar product $\langle f, g \rangle = \sum_{x \in \mathscr{X}_V} f(x)\, g(x)$ in $\mathbb{R}^{\mathscr{X}_V}$. The definition of pure interaction spaces will be used for the derivation of the dimension formulas (3) and (4) and will play no further role in this paper. One has the following orthogonal decomposition of spaces of A-interactions into pure interactions:

$$\mathcal{I}_A = \bigoplus_{B \subseteq A} \widetilde{\mathcal{I}}_B.$$

The symbol "$\bigoplus$" denotes the direct sum of orthogonal vector spaces, and therefore $\mathcal{I}_A$ is the smallest subspace of $\mathscr{X}^{\mathbb{R}_V}$ that contains all the $\widetilde{\mathcal{I}}_B$, $B \subseteq A$. This implies

$$\mathsf{dim}\,(\mathcal{I}_A) \;=\; \sum_{B \subseteq A} \mathsf{dim}\left(\widetilde{\mathcal{I}}_B\right),$$

and with the Möbius inversion formula we obtain

$$\mathsf{dim}(\widetilde{\mathcal{I}}_A) \;=\; \sum_{B\subseteq A}(-1)^{|A\setminus B|}\prod_{v\in B}|\mathscr{X}_v| \;=\; \prod_{v\in A}(|\mathscr{X}_v|-1)\,.$$

Given a simplicial complex $\mathfrak{A}$, we consider the sum

$$\mathcal{I}_{\mathfrak{A}} \;:=\; \sum_{A\in\mathfrak{A}}\mathcal{I}_A \;=\; \bigoplus_{A\in\mathfrak{A}}\widetilde{\mathcal{I}}_A$$

which has dimension

$$\mathsf{dim}(\mathcal{I}_{\mathfrak{A}}) \;=\; \sum_{A\in\mathfrak{A}}\prod_{v\in A}(|\mathscr{X}_v|-1). \tag{3}$$

In the simple example of binary nodes, that is $|\mathscr{X}_v|=2$ for all v, we get $|\mathfrak{A}|$ as dimension. We define a *hierarchical model* as

$$\mathcal{E}_{\mathfrak{A}} \;:=\; \left\{\frac{e^f}{\sum_x e^{f(x)}} \;:\; f\in\mathcal{I}_{\mathfrak{A}}\right\} \;\subseteq\; \mathcal{P}(\mathscr{X}_V),$$

where $\mathcal{P}(\mathscr{X}_V)$ denotes the set of strictly positive probability measures on $\mathscr{X}_V$. It is a subset of $\mathbb{R}^{\mathscr{X}_V}$ and carries the naturally induced topology. Throughout the paper, the topological closure of a set $\mathcal{E}\subseteq\mathbb{R}^{\mathscr{X}_V}$ will be denoted by $\overline{\mathcal{E}}$.

The dimension of $\mathcal{E}_{\mathfrak{A}}$ is one less than the dimension of $\mathcal{I}_{\mathfrak{A}}$:

$$\mathsf{dim}\big(\mathcal{E}_{\mathfrak{A}}\big) \;=\; \sum_{\substack{A\in\mathfrak{A}\\ A\neq\emptyset}}\prod_{v\in A}(|\mathscr{X}_v|-1)\,. \tag{4}$$

Here, we have used the convention $\prod_{v\in A}f_v=1$ if $A=\emptyset$.

Example 2.2. Of particular interest are the simplicial complexes that are controlled by the maximal interaction order within the system:

$$\mathfrak{A}_k \;:=\; \bigcup_{l=0}^{k}\binom{V}{l}, \qquad k=1,2,\ldots,N=|V|.$$

Here, $\binom{V}{l}$ denotes the set subsets of V that have l elements. The corresponding hierarchical models $\mathcal{E}^{(k)}$ consist of strictly positive distributions that can be described by interactions up to order k and have dimension

$$\mathsf{dim}\left(\mathcal{E}^{(k)}\right) \;=\; \sum_{l=1}^{k}\binom{N}{l}.$$

★

It is well known within information geometry that the closure of a hierarchical model $\mathcal{E}_{\mathfrak{A}}$ coincides with the image of the maximum entropy projection P described in Section 2.2.1. Furthermore, the maximum entropy estimate coincides with the unique probability measure $\pi_{\mathfrak{A}}(p)$ in the topological closure of $\mathcal{E}_{\mathfrak{A}}$ satisfying

$$D(p\,\|\,\pi_{\mathfrak{A}}(p)) \;=\; \inf_{q\in\mathcal{E}_{\mathfrak{A}}} D(p\,\|\,q) \;=:\; D(p\,\|\,\mathcal{E}_{\mathfrak{A}}), \tag{5}$$

and we have

$$D(p\,\|\,\pi_{\mathfrak{A}}(p)) \;=\; H_{\pi_{\mathfrak{A}}(p)}(X_V)-H_p(X_V). \tag{6}$$

(See for instance the general treatment [14].) Following our general geometric approach, the deviation $D(p\,\|\,\mathcal{E}_{\mathfrak{A}})$ from $\mathcal{E}_{\mathfrak{A}}$ is interpreted as a complexity measure. For the exponential families $\mathcal{E}^{(k)}$ of

Example 2.2 this divergence quantifies the extent to which p cannot be explained in terms of interactions of maximal order k. Stated differently, it quantifies the extent to which the whole is more than the sum of its parts of size k. For $k = 1$, we refer to the quantity $D(p \,\|\, \mathcal{E}^{(1)})$ as *multi-information*. In order to define multi-information more generally, consider a partition $\xi = \{A_1, \ldots, A_n\}$ of the node set V and the corresponding simplicial complex $\mathfrak{A} = \bigcup_{i=1}^n 2^{A_i}$. In this special case, the measure $D(p \,\|\, \mathcal{E}_{\mathfrak{A}})$ can be calculated explicitly, and we have the following entropic representation:

$$\begin{aligned} I(X_{A_1}; \ldots; X_{A_n}) &:= D(p \,\|\, \mathcal{E}_{\mathfrak{A}}) \\ &= \sum_{i=1}^n H(X_{A_i}) - H(X_{A_1}, \ldots, X_{A_n}). \end{aligned}$$

In the next section we study the maximization of $D(\cdot \,\|\, \mathcal{E}_{\mathfrak{A}})$.

2.2.3. *The maximization of complexity.* We now wish to discuss whether the maximization of $D(\cdot \,\|\, \mathcal{E}_{\mathfrak{A}})$ should be considered as a good approach for finding systems of maximal complexity. We shall state a result about restrictions on the structure of maximizers of this functional. Interpreting this result in a positive way leads us to the conclusion that the rules can be revealed that underly complex systems and thereby allow for understanding and controlling complexity. However, we then also have to face the somewhat paradoxical fact that the most complex systems, according to this criterion, are in a certain sense rather simple. More precisely, they have a strongly restricted support set, as shown in the following basic result that follows from general results [2], [28], [29]:

Proposition 2.3. *Let $\mathfrak{A}$ be a simplicial complex, and let p be a maximizer of $D(\cdot \,\|\, \mathcal{E}_{\mathfrak{A}})$. Then*

$$|\mathsf{supp}(p)| \;\leq\; \mathsf{dim}(\mathcal{E}_{\mathfrak{A}}) + 1 \;=\; \sum_{A \in \mathfrak{A}} \prod_{v \in A} (|\mathscr{X}_v| - 1). \tag{7}$$

Furthermore,

$$p(x) \;=\; \frac{\pi_{\mathfrak{A}}(p)(x)}{\sum_{x' \in \mathsf{supp}(p)} \pi_{\mathfrak{A}}(p)(x')}, \qquad x \in \mathsf{supp}(p).$$

Stated informally, this result shows that the maximizers of complexity have a reduced support and that they coincide with their projection, interpreted as non-complex systems, on this support. For binary nodes, the support reduction is quite strong. In that case, the inequality (7) becomes $|\mathsf{supp}(p)| \leq |\mathfrak{A}|$. Figure 2 illustrates these simplicity aspects of the maximizers of complexity.

The above-mentioned simplicity aspects of complex systems become more explicit in the particular context of the hierarchical models $\mathcal{E}^{(k)}$ introduced in Example 2.2. We ask the following question:

Which order m of interaction is sufficient for generating all distributions that (locally) maximize the deviation from being the sum of parts of size k?

For $k = 1$, this question has been addressed in [4]. It turns out that, if the cardinalities $|\mathscr{X}_v|$, $v \in V$, satisfy a particular combinatorial condition, the interaction order $m = 2$ is sufficient for generating the global maximizers of the multi-information. More precisely, if p is a global maximizer of the multi-information then p is contained in the closure of $\mathcal{E}^{(2)}$. These are the probability measures that can be completely described in terms of pairwise interactions, and they include the probability measures corresponding to complete synchronization. In the case of two binary nodes, the maximizers are the measures $p_+ := \frac{1}{2}\left(\delta_{(0,0)} + \delta_{(1,1)}\right)$ and $p_- := \frac{1}{2}\left(\delta_{(1,0)} + \delta_{(0,1)}\right)$ in Figure 2. Note, however, that in this low-dimensional case the closure of $\mathcal{E}^{(2)}$ coincides with the whole simplex

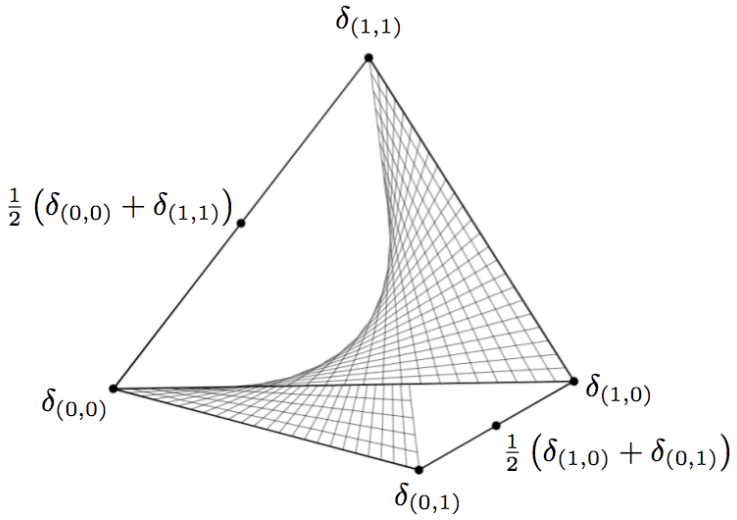

FIGURE 2. The three-dimensional simplex and its two-dimensional subfamily of product distributions. The extreme points of the simplex are the Dirac measures $\delta_{(x,y)}$, $x, y \in \{0,1\}$. The maximization of the distance from the family of product distributions leads to distributions with support cardinality two. In addition, the maximizers have a very simple structure.

of probability measures on $\{0,1\} \times \{0,1\}$. Therefore, being contained in the closure of $\mathcal{E}^{(2)}$ is a property of all probability measures and not special at all. The situation changes with three binary nodes, where the dimension of the simplex $\mathcal{P}(\{0,1\}^3)$ is seven and the dimension of $\mathcal{E}^{(2)}$ is six. For general k, the above question can be addressed using the following result [22]:

$$|\mathsf{supp}(p)| \leq s, \;\; m \geq \log_2(s+1) \qquad \Rightarrow \qquad p \in \overline{\mathcal{E}^{(m)}}. \tag{8}$$

According to Proposition 2.3 the maximizers of $D(\cdot \,\|\, \mathcal{E}^{(k)})$ have support of size smaller than or equal to $\sum_{l=0}^{k} \binom{N}{l}$. If $m \geq \log_2\left(1 + \sum_{l=0}^{k} \binom{N}{l}\right)$ then, according to (8), all these maximizers will be in the closure of $\mathcal{E}^{(m)}$. For example, if we have one thousand binary nodes, that is $N = 1000$, and maximize the multi-information $D(\cdot \,\|\, \mathcal{E}^{(1)})$ then interaction of order ten is sufficient for generating all distributions with locally maximal multi-information. This means that a system of size one thousand with (locally) maximal deviation from the sum of its elements (parts of size one) is not more than the sum of its parts of size ten. In view of our understanding of complexity as the extent to which the system is more than the sum of its parts of any size, it appears inconsistent to consider these maximizers of multi-information as complex. One would assume that complexity is reflected in terms of interactions up to the highest order (see reference [24], which analyzes coupled map lattices and cellular automata from this perspective). Trying to resolve this apparent inconsistency by maximizing the distance $D(\cdot \,\|\, \mathcal{E}^{(2)})$ from the larger exponential family $\mathcal{E}^{(2)}$ instead of maximizing the multi-information $D(\cdot \,\|\, \mathcal{E}^{(1)})$ does not lead very far. We can repeat the argument and observe that interaction of order nineteen is now sufficient for generating all the corresponding maximizers: A system of size one thousand with maximal deviation from the sum of its parts of size two is not more than the sum of its parts of size nineteen.

In this paper, we propose a way to address this inconsistency and draw connections to other approaches to complexity. We discuss two examples, the TSE complexity and excess entropy.

Consider a hierarchy of simplicial complexes

$$\mathfrak{A}_1 \subseteq \mathfrak{A}_2 \subseteq \ldots \subseteq \mathfrak{A}_{N-1} \subseteq \mathfrak{A}_N := 2^V$$

and denote the projection $\pi_{\mathfrak{A}_k}$ on $\mathcal{E}_{\mathfrak{A}_k}$ by $p^{(k)}$. Then the following equality holds:

$$D(p \,\|\, p^{(1)}) = \sum_{k=1}^{N-1} D(p^{(k+1)} \,\|\, p^{(k)}). \tag{9}$$

We shall use the same notation $p^{(k)}$ for various hierarchies of simplicial complexes. Although being clear from the given context, the particular meaning of these distributions will change throughout the paper.

The above considerations imply that when the left hand side of (9) is maximized, only the first few terms on the right hand side are dominant and remain positive. Therefore, instead of (9) we consider a weighted sum with a weight vector $\alpha = (\alpha(1), \ldots, \alpha(N-1)) \in \mathbb{R}^{(N-1)}$ and set:

$$C_\alpha(p) := \sum_{k=1}^{N-1} \alpha(k)\, D(p \,\|\, p^{(k)}) \tag{10}$$

$$= \sum_{k=1}^{N-1} \beta(k)\, D(p^{(k+1)} \,\|\, p^{(k)}) \tag{11}$$

with $\beta(k) := \sum_{i=1}^{k} \alpha(i)$. As we have seen, the multi-information can be represented in this way by setting $\alpha(1) := 1$, and $\alpha(k) := 0$ for $k \geq 2$, or, equivalently, $\beta(k) = 1$ for all k. This makes clear that our ansatz provides a general structure, in place of specifying a distinguished complexity measure. If one wants to specify such a measure, one has to identify the correct weight vector α by means of additional assumptions. Generating complex systems would then require forcing *all* contributions $D(p^{(k+1)} \,\|\, p^{(k)})$ to display a specific shape of behaviour as k increases. In order to have at least positivity of these contributions for maximal C_α, it appears reasonable to assume that the sequence $\beta(k)$ is strictly increasing with k. This is the case if and only if all the $\alpha(k)$ are positive. Clearly, multi-information does not satisfy this assumption. In the next section we will introduce TSE complexity and excess entropy and show how they fit into our framework with reasonable weight vectors.

3. Comparison with other approaches

3.1. **TSE-Complexity.** As we have already demonstrated in previous sections, although multi-information perfectly fits into the concept of complexity as deviation of the whole from the sum of its parts of size one, its maximization leads to distributions that do not appear very complex. In particular, multi-information is maximized for distributions with complete synchrony of the nodes' states. Tononi, Sporns, and Edelman [37] introduced a measure of brain complexity which incorporates multi-information as integration capacity of the underlying system. In addition to this capacity they also require the differentiation capacity as necessary contribution to the system's complexity. The interplay between integration and differentiation then generates distributions of more "complex" global configurations where distributions with total synchronization are identified as simple. In order to introduce their complexity measure, consider for each $1 \leq k \leq N-1$ the

quantity

$$C_p^{(k)} := I_p(X_1;\ldots;X_N) - \frac{N}{k\binom{N}{k}} \sum_{1\le i_1<\cdots<i_k\le N} I_p(X_{i_1};\ldots;X_{i_k}) \tag{12}$$

$$= \frac{N}{k\binom{N}{k}} \sum_{1\le i_1<\cdots<i_k\le N} H_p(X_{i_1},\ldots,X_{i_k}) - H_p(X_1,\ldots,X_N) \tag{13}$$

$$= \frac{N}{k} H_p(k) - H_p(N) \tag{14}$$

where we use the abbreviation

$$H_p(k) := \frac{1}{\binom{N}{k}} \sum_{1\le i_1<\cdots<i_k\le N} H_p(X_{i_1},\ldots,X_{i_k})$$

to denote the average entropy of subsets of size k. The *TSE-complexity* is then defined as a weighted sum of the

$$C_p := \sum_{k=1}^{N-1} \frac{k}{N} C_p^{(k)}.$$

This corresponds to the general structure (10) with weights $\alpha(k) = \frac{k}{N}$.

The $C_p^{(k)}$'s quantify the deviation of the mean stochastic dependence of subsets of size k from the total stochastic dependence. Therefore, at least conceptually, they correspond to the distances $D(p \,\|\, p^{(k)})$ where $p^{(k)}$ denotes the maximum entropy estimate that has the same k'th order marginals as p. We will see in Corollary 3.5 that the $C_p^{(k)}$ are monotone as well. The following theorem summarizes the main result of this section, further relating the $C_p^{(k)}$ to the $D(p \,\|\, p^{(k)})$.

Theorem 3.1. *For all k, $1 \le k \le N$, we have*

$$D(p \,\|\, p^{(k)}) \le C_p^{(k)}.$$

This states that the $C_p^{(k)}$ can be considered as an upper estimate of those dependencies that can not be described in terms of interactions up to order k. Theorem 3.1 follows from properties of the average subset entropies. They are of interest in their own right and therefore we present them in what follows. The proofs are given in the appendix. First, we show that the difference between $H_p(k)$ and $H_p(k-1)$ can be expressed as an average over conditional entropies. In what follows, we will simplify notation and neglect the subscript p whenever it is appropriate.

Proposition 3.2.

$$H(k) - H(k-1) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\binom{N-1}{k-1}} \sum_{\substack{1\le i_1<\cdots<i_{k-1}\le N \\ i\neq i_1,\ldots,i_{k-1}}} H(X_i | X_{i_1},\ldots,X_{i_{k-1}}) =: h(k). \tag{15}$$

The next step is then to show that differences between the averaged conditional entropies $h(k)$ and $h(k+1)$ can be expressed as an average over conditional mutual informations.

Proposition 3.3.

$$h(k) - h(k+1) \;=\; \frac{1}{N(N-1)} \sum_{\substack{1 \le i,j \le N \\ i \neq j}} \frac{1}{\binom{N-2}{k-1}} \sum_{\substack{1 \le i_1 < \cdots < i_{k-1} \le N \\ i,j \neq i_1,\ldots,i_{k-1}}} I(X_i; X_j | X_{i_1}, \ldots, X_{i_{k-1}}).$$

Since conditional mutual informations are positive, we can conclude that $h(k+1) \le h(k)$, i.e. the $H(k)$ form a monotone and concave sequence as shown in Figure 3.

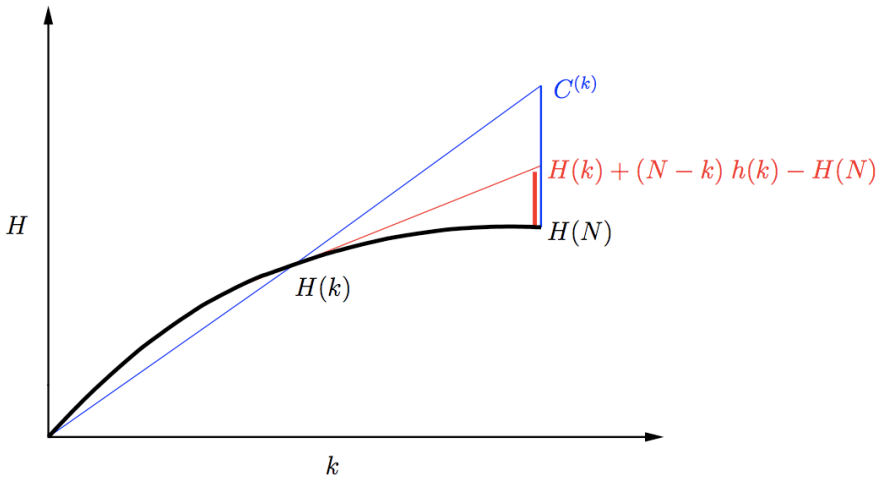

FIGURE 3. The average entropy of subsets of size k grows with k. $C^{(k)}$ can be considered to be an estimate of the system entropy $H(N)$ based on the assumption of a linear grows through $H(k)$.

Furthermore, $H(k)$ can be expressed as the sum of differences

$$H(k) = \sum_{i=1}^{k} h(i)$$

with the convention $H(0) = 0$. Using Proposition 3.3 again, we can prove the following

Lemma 3.4.

$$H(k) \;\ge\; k\, h(k) \;\ge\; k\, h(k+1).$$

This in turn allows to show that the terms of the TSE complexity are monotone.

Corollary 3.5.

$$C^{(k)} \;\le\; C^{(k-1)}.$$

Another consequence of Propositions 3.2 and 3.3 is that the entropy $H_{p^{(k)}}(X_V)$ of the maximum entropy distribution can be bounded using the entropy of the marginals of size k:

$$H_{p^{(k)}}(X_V) \;\le\; H_p(k) + (N-k)\, h_p(k).$$

This also provides an upper bound for the distance

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

as

$$D(p \,\|\, p^{(k)}) \;\leq\; H_p(k) + (N-k)\; h_p(k) - H_p(N). \tag{17}$$

Similarly, by (14), the factors of the TSE-complexity can be expressed as

$$\begin{aligned} C_p^{(k)} &= \frac{N}{k} H_p(k) - H_p(N) \\ &= H_p(k) + (N-k)\; \frac{1}{k} H_p(k) - H_p(N). \end{aligned} \tag{18}$$

We may therefore consider $C_p^{(k)}$ as an approximation of $D(p \,\|\, p^{(k)})$ that uses $\frac{1}{k} H_p(k)$ instead of $h_p(k)$ in order to extrapolate the system entropy from the entropy of marginals of size k as illustrated in Figure 3. We have already shown in Lemma 3.4 that the estimate (18) is less precise than (17), i.e. $\frac{1}{k} H_p(k) \geq h_p(k)$, and thus obtain

$$D(p \,\|\, p^{(k)}) \;\leq\; C_p^{(k)},$$

which proves Theorem 3.1.

3.2. **Excess entropy.** Similar to $C^{(1)}$ — the multi-information, which appears in the first summand of the TSE-complexity — also $C^{(N-1)}$, which occurs in the last summand, is related to a complexity measure of its own: The *excess entropy* for finite systems:

$$E_p(X_V) \;=\; H_p(X_V) - \sum_{i \in V} H_p(X_i | X_{V \setminus i}). \tag{19}$$

Here one subtracts from the uncertainty about the state of the whole system the remaining uncertainties of the states of the elements given the state of the other elements. This complexity measure was described as *dual total correlation* already in [20]. For a more detailed discussion of its properties see [9, 30]. In the present context it is important that, by (14, 15), it can be expressed as

$$\begin{aligned} E_p(X_V) &= N H_p(N-1) - (N-1) H_p(N) \\ &= (N-1)\, C_p^{(N-1)}. \end{aligned}$$

Therefore it follows from Theorem 3.1 that the distance $D(p \,\|\, p^{(N-1)})$ provides a lower bound for this complexity measure

$$(N-1) D(p \,\|\, p^{(N-1)}) \;\leq\; E_p(X_V). \tag{20}$$

The term *excess entropy* is used more frequently in the context of stochastic processes [14]. In this case the excess entropy is a very natural complexity measure because it measures the amount of information that it nessecary to perform an optimal prediction. It is also known as *effective measure complexity* [19] and as *predictive information* [10]. In a stochastic process X, the set of nodes V exhibits a temporal order $X_1, X_2, \ldots, X_N, \ldots$, and in what follows we assume that the distribution of this sequence is invariant with respect to the shift map $(x_1, x_2, \ldots) \mapsto (x_2, x_3, \ldots)$. The uncertainty of a single observation X_N is given by the marginal entropy $H(X_N)$. The uncertainty of this observation when the past $N-1$ values are known is quantified by

$$h_N \;:=\; H(X_N | X_1, \ldots, X_{N-1})$$

with the limit, if it exists,

$$h_\infty \;:=\; \lim_{N\to\infty} h_N \tag{21}$$

called the entropy rate of the process. The excess entropy of the process with the entropy rate h_∞ is then

$$E(X) \;:=\; \lim_{N\to\infty} \left(H(X_1,\ldots,X_N) - Nh_\infty\right) \tag{22}$$

It measures the nonextensive part of the entropy, i.e. the amount of entropy of each element that *exceeds* the entropy rate. In what follows we shall derive a representation of the excess entropy $E(X)$ in terms of (10). This will be done in several steps. All corresponding proofs are provided in the appendix.

For each $N \in \mathbb{N}$ we consider the probability distribution $p \in \overline{\mathcal{P}(\mathscr{X}^N)}$ defined by

$$p_N(x_1,\ldots,x_N) \;:=\; \mathsf{Prob}\left\{X_1 = x_1, \ldots, X_N = x_N\right\}, \qquad x_1,\ldots,x_N \in \mathscr{X}\,.$$

In the following we use the interval notation $[r,s] = \{r, r+1, \ldots, s\}$ and $X_{[r,s]} = x_{[r,s]}$ for $X_r = x_r$, $X_{r+1} = x_{r+1}$, …, $X_s = x_s$. We consider the family of simplicial complexes

$$\mathfrak{A}_{N,k+1} \;:=\; \{[r,s] \subseteq [1,N] \;:\; s - r \leq k\}, \qquad 0 \leq k \leq N-1\,.$$

The corresponding hierarchical models $\mathcal{E}_{\mathfrak{A}_{N,k+1}} \subseteq \mathcal{P}(\mathscr{X}^N)$ represent the Markov processes of order k. As the following proposition shows, the maximum entropy projection coincides with the k-order Markov approximation of the process $X_{[1,N]}$.

Proposition 3.6. *Let $X_1, X_2, \ldots, X_N$ be random variables in a non-empty and finite state set $\mathscr{X}$ with joint probability vector $p \in \overline{\mathcal{P}(\mathscr{X}^N)}$, and let $p^{(k)}$ denote the maximum entropy estimate of p with respect to $\mathfrak{A}_{N,k+1}$. Then*

$$p^{(k+1)}(x_1,\ldots,x_N) \;=\; p(x_1,\ldots,x_{k+1}) \prod_{i=2}^{N-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}), \tag{23}$$

$$D(p \,\|\, p^{(k+1)}) \;=\; \sum_{i=1}^{N-k-1} I_p(X_{[1,i]}; X_{k+i+1} \,|\, X_{[i+1,k+i]}). \tag{24}$$

We have the following special cases of (23):

$$\begin{aligned} p^{(1)}(x_1,\ldots,x_N) &= \prod_{i=1}^{N} p(x_i), \\ p^{(2)}(x_1,\ldots,x_N) &= p(x_1)\,p(x_2|x_1)\cdots p(x_N|x_{N-1}), \\ p^{(N)}(x_1,\ldots,x_N) &= p(x_1,\ldots,x_N). \end{aligned}$$

Proposition 3.7. *In the situation of Proposition 3.6 we have*

$$D(p^{(k+1)} \,\|\, p^{(k)}) \;=\; \sum_{i=1}^{N-k} I_p(X_{k+i}; X_i \,|\, X_{[i+1,k+i-1]}), \qquad 1 \leq k \leq N-1. \tag{25}$$

If the process is stationary, the right hand side of (25) equals $(N-k)\, I_p(X_1; X_{k+1} \,|\, X_{[2,k]})$.

Given a stochastic process $X = (X_k)_{k\in\mathbb{N}}$ with non-empty and finite state set $\mathscr{X}$, one has the following alternative representation of the excess entropy [14]:

$$E(X) \;=\; \sum_{k=1}^{\infty} k\, I(X_1; X_{k+1} \,|\, X_{[2,k]}). \tag{26}$$

We apply Proposition 3.7 to this representation and obtain

$$\begin{aligned} E_p(X) &= \sum_{k=1}^{\infty} k\, I_p(X_1; X_{k+1}|X_{[2,k]}) \;=\; \lim_{N\to\infty} \sum_{k=1}^{N-1} k\, I_p(X_1; X_{k+1}|X_{[2,k]}) \\ &= \lim_{N\to\infty} \sum_{k=1}^{N-1} \frac{k}{N-k}\,(N-k)\, I_p(X_1; X_{k+1}|X_{[2,k]}) \\ &= \lim_{N\to\infty} \underbrace{\sum_{k=1}^{N-1} \frac{k}{N-k}\, D(p_N{}^{(k+1)} \,\|\, p_N{}^{(k)})}_{=:E_N} \end{aligned}$$

Thus, we have finally obtained a representation of quantities E_N that have the structure (11) and converge to the excess entropy. The corresponding weights $\beta(k) = \frac{k}{N-k}$ are strictly increasing with k. Note that in the context of time series, the projections $p_N^{(k)}$ are defined with respect to other exponential families than in the previous section. The essential difference is that in the time series context we have a complete order of the nodes so that we can define a simplicial complex in terms of intervals of a paricular length, say length k. In the general case, where we do not have such an ordering, we define the corresponding simplicial complex in terms of *all* subsets of size k. This clearly generates a larger hierarchical model and therefore a smaller distance. Furthermore, the special case of intervals allows us to compute the projections onto the hierarchical models and the corresponding distances explicitly. It turns out that these distances are nothing but conditional mutual informations [23]. Note that such explicit calculations are in general not possible.

4. Conclusions

Complexity is considered as emerging from interactions between elements, or, better and more generally, parts of a system. When formalizing this in terms of information-theoretic quantities, one is led to interactions of random variables. We have carried out such a formalization for finite systems. In order to analyze interactions, we implement the idea of decomposing the stochastic dependence among the parts of a given system. Such a decomposition needs to go beyond representations of stochastic dependence in terms of marginal entropies (see (13) as an example). For our more general analysis, information geometry provides the natural framework of hierarchies of exponential families that makes an "orthogonal" decomposition of the underlying joint distribution possible with respect to the interaction order ([1]). While well-known complexity measures such as the TSE-complexity or the excess entropy are defined in terms of marginal entropies we propose the following general family of complexity measures with free parameters $\alpha(k)$:

$$C_\alpha(p) \;:=\; \sum_{k=1}^{N-1} \alpha(k)\, D(p \,\|\, p^{(k)}) \;=\; \sum_{k=1}^{N-1} \beta(k)\, D(p^{(k+1)} \,\|\, p^{(k)}),$$

where $\beta(k) := \sum_{i=1}^{k} \alpha(i)$. This family reflects the "orthogonal" decomposition of the stochastic dependence into contributions corresponding to a hierarchy of interactions among the constituents.

The ansatz C_α is very general and incorporates many known complexity measures as special cases. We show that, in particular, the TSE complexity and the excess entropy can be captured by our approach by appropriately choosing the weights. In the case of TSE complexity we identify the weights $\alpha(k) = \frac{k}{N}$ which correspond to $\beta(k) = \frac{k\,(k+1)}{2}$. The excess entropy can be expressed with weights $\beta(k) = \frac{k}{N-k}$. In both cases, the sequence $\beta(k)$ increases with k. In contrast, the weights of the multi-information have a "flat" shape: $\beta(k) = 1$ for all k. We demonstrate the fact that this shape leads to the somewhat paradoxical situation in which complex systems can be understood in terms of their parts of a particular size that is not maximal. We argue that, in order to avoid this situation, one needs to "strengthen" the higher order contributions by a sequence $\beta(k)$ that increases with k. In this sense, the weights of the TSE complexity and the excess entropy lead to reasonable complexity measures.
An interesting line for further research will be the identification of (other) rationales to choose the weighting factors $\alpha(k)$ that could provide new insight- and useful complexity measures. In this regard, complexity measures with weights that normalize the individual contributions to C_α are of particular interest. Furthermore, natural assumptions on the scaling properties of complexity could lead to a structural determination of the weights. Related studies of TSE complexity and excess entropy can be helpful in this context [30].

Acknowledgements

This work has been supported by the Santa Fe Institute and the Volkswagen Foundation.

References

[1] S. Amari. *Information geometry on hierarchy of probability distributions.* IEEE Trans. IT **47** (2001) 1701–1711.
[2] N. Ay. *An information-geometric approach to a theory of pragmatic structuring.* Ann. Prob. **30** (2002) 416–436.
[3] N. Ay. *Information geometry on complexity and stochastic interaction.* MPI MIS Preprint 95/2001.
[4] N. Ay, A. Knauf. *Maximizing Multi-Information.* Kybernetika 42 (5) (2007) 517–538.
[5] N. Ay, M. Müller, A. Szkoła. *Effective Complexity and its Relation to Logical Depth.* IEEE Transactions on Information Theory 56 (9) (2010) 4593–4607.
[6] N. Ay, M. Müller, A. Szkoła. *Effective Complexity of Stationary Process Realizations.* Entropy 13 (2011) 1200–1211.
[7] S. Amari, H. Nagaoka. *Methods of Information Geometry.* Oxford University Press 2000.
[8] N. Ay, T. Wennekers. *Dynamical Properties of Strongly Interacting Markov Chains.* Neural Networks 16 (2003) 1483–1497.
[9] N. Ay, E. Olbrich, N. Bertschinger, J. Jost. *A unifying framework for complexity measures of finite systems.* Proceedings of ECCS'06 (2006). Santa Fe Institute Working Paper 06-08-028.
[10] W. Bialek, I. Nemenman, N. Tishby. *Predictability, Complexity, and Learning.* Neural Computation 13 (2001) 2409–2463.
[11] G. J. Chaitin. *On the Length of Programs for Computing Binary Sequences.* J. Assoc. Comp. Mach. 13 (1966) 547–569.
[12] T. Cover, J. Thomas, *Elements of Information Theory.* Wiley 1991.
[13] J. P. Crutchfield and K. Young. *Inferring Statistical Complexity.* Phys. Rev. Lett. 63 (1989) 105–108.
[14] J. P. Crutchfield and David P. Feldman. *Regularities unseen, randomness observed: Levels of entropy convergence.* Chaos 13 (1) (2003) 25–54.
[15] I. Csiszár and F. Matúš. *Information projections revisited.* IEEE Transactions Information Theory 49 (2003) 1474–1490.

[16] D. P. Feldman, J. P. Crutchfield. *Synchronizing to Periodicity: The Transient information and Synchronization Time of Periodic Sequences.* Adv. Compl. Sys. 7 (2004) 329–355.
[17] M. Gell-Mann, S. Lloyd. *Information Measures, Effective Complexity, and Total Information.* Complexity 2 (1996) 44–52.
[18] M. Gell-Mann, S. Lloyd. *Effective Complexity.* Santa Fe Institute Working Paper 03-12-068 (2003).
[19] P. Grassberger. *Toward a quantitative theory of self-generated complexity.* Int. J. Theor. Phys. 25 (9) (1986) 907–938.
[20] T. S. Han. *Nonnegative Entropy Measures of Multivariate Symmetric Correlations.* Information and Control 36 (1978) 133–156.
[21] J. Hopcroft, R. Motvani, J. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison Wesley Longman ^{2}2001.
[22] T. Kahle. *Neighborliness of Marginal Polytopes.* Contributions to Algebra and Geometry 51(1) (2010) 45–56.
[23] K. Lindgren. *Correlations and Random Information in Cellular Automata.* Complex Systems 1 (1987) 527–543.
[24] T. Kahle, E. Olbrich, J. Jost, N. Ay. *Complexity Measures from Interaction Structures.* Phys. Rev. E 79 (2009) 026201.
[25] A. N. Kolmogorov. *Three Approaches to the Quntitative Definition on Information.* Problems of Information Transmission 1 (1965) 4–7.
[26] A. N. Kolmogorov. *A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces.* (Russian) Dokl. Akad. Nauk SSSR (N.S.) 119 (1958) 861–864.
[27] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer 1997.
[28] F. Matúš, N. Ay. *On maximization of the information divergence from an exponential family.* Proceedings of WUPES'03 (ed. J. Vejnarova), University of Economics Prague (2003) 199–204.
[29] F. Matúš. *Optimality conditions for maximizers of the information divergence from an exponential family.* Kybernetika 43 (2007) 731–746.
[30] E. Olbrich, N. Bertschinger, N. Ay, J. Jost. *How should complexity scale with system size?* Eur. Phys. J. B 63 (2008) 407–415.
[31] C. Papadimitriou. *Computational Complexity.* Addison Wesley 1994.
[32] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific 1989.
[33] C. E. Shannon. *A mathematical theory of communication.* Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.
[34] Ja. Sinai *On the concept of entropy for a dynamical system.* (Russian) Dokl. Akad. Nauk SSSR 124 (1959) 768–771.
[35] C. R. Shalizi, J. P. Crutchfield. *Computational mechanics: Pattern and prediction, structure and simplicity.* Journal of Statistical Physics 104 (2001) 817–879.
[36] R. J. Solomonoff. *A Formal Theory of Inductive Inference.* Inform. Contr. 7 (1964) 1–22, 224–254.
[37] G. Tononi, O. Sporns, G. M. Edelman. *A measure for brain complexity: Relating functional segregation and integration in the nervous systems.* Proc. Natl. Acad. Sci. USA (91) (1994) 5033–5037.

5. Appendix: Technical prerequisites and proofs of the results

5.1. **Maximum entropy estimation in hierarchical models.**

Lemma 5.1. *Let $\mathfrak{A}$ be a simplicial complex and let p be a distribution in the closure of $\mathcal{P}(\mathscr{X}_V)$ (not necessarily positive). Then: If a distribution p^* satisfies the following two conditions then it is the maximum entropy estimate of p with respect to $\mathfrak{A}$:*

(1) *There exist functions $\phi_A \in \mathcal{I}_A$, $A \in \mathfrak{A}$, satisfying $p^*(x) = \prod_{A\in\mathfrak{A}} \phi_A(x_A)$, and*
(2) *for all $A \in \mathfrak{A}$ the A-marginal of p^* coincides with the A-marginal of p, that is ${p^*}_A = p_A$.*

Proof. Let q be a probability measure that satisfies $q_A = p_A$ for all $A \in \mathfrak{A}$. We prove that the entropy of p^* is greater than or equal to the entropy of q in three steps.
Step 1: $\mathsf{supp}(q) \subseteq \mathsf{supp}(p^*)$: Let $x \in \mathsf{supp}(q)$. Then for all $B \in \mathfrak{A}$ one has

$$0 \;<\; q(x) \;\le\; q_B(x_B) \;=\; {p^*}_B(x_B) \;=\; \sum_{x_{V\setminus B}} \prod_{A\in\mathfrak{A}} \phi_A(x_A) \;=\; \phi_B(x_B) \sum_{x_{V\setminus B}} \prod_{\substack{A\in\mathfrak{A}\\ A\neq B}} \phi_A(x_A)$$

and therefore $\phi_B(x_B) > 0$. This implies $p^*(x) > 0$.
Step 2: It is easy to see that

$$q_A \;=\; {p^*}_A \text{ for all } A \in \mathfrak{A} \qquad \Leftrightarrow \qquad q(f) \;=\; p^*(f) \text{ for all } f \in \mathcal{I}_{\mathfrak{A}}. \tag{27}$$

In Step 3 we apply (27) to a particular function in $\mathcal{I}_{\mathfrak{A}}$. To this end, we define $\widetilde{\phi}_A \in \mathcal{I}_A$ by $\widetilde{\phi}_A(x_A) := 1$ if $\phi_A(x_A) = 0$ and $\widetilde{\phi}_A(x_A) := \phi_A(x_A)$ otherwise and consider

$$\sum_{A\in\mathfrak{A}} \log_2 \widetilde{\phi}_A \;\in\; \mathcal{I}_{\mathfrak{A}}. \tag{28}$$

Step 3:

$$\begin{aligned}
H_{p^*}(X_V) &= -\sum_{x\in\mathsf{supp}(p^*)} p^*(x) \log_2 p^*(x) \\
&= -\sum_{x\in\mathsf{supp}(p^*)} p^*(x) \sum_{A\in\mathfrak{A}} \log_2 \phi_A(x_A) \\
&= -\sum_{x} p^*(x) \sum_{A\in\mathfrak{A}} \log_2 \widetilde{\phi}_A(x_A) \\
&= -\sum_{x} q(x) \sum_{A\in\mathfrak{A}} \log_2 \widetilde{\phi}_A(x_A) \qquad \text{(application of (27) to the function (28))} \\
&= -\sum_{x\in\mathsf{supp}(q)} q(x) \sum_{A\in\mathfrak{A}} \log_2 \widetilde{\phi}_A(x_A) \\
&= -\sum_{x\in\mathsf{supp}(p^*)} q(x) \sum_{A\in\mathfrak{A}} \log_2 \phi_A(x_A) \qquad \text{(Step 1)} \\
&= -\sum_{x\in\mathsf{supp}(p^*)} q(x) \log_2 p^*(x) \qquad \text{(cross entropy)} \\
&\ge H_q(X_V).
\end{aligned}$$

■

5.2. **Proofs of the main results of the paper.**

Proof. (Proposition 3.2)

$$\begin{aligned}
h(k) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\binom{N-1}{k-1}}\sum_{\substack{1\le i_1<\cdots<i_{k-1}\le N\\ i\neq i_1,\ldots,i_{k-1}}} H(X_i|X_{i_1},\ldots,X_{i_{k-1}})\\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\binom{N-1}{k-1}}\sum_{\substack{1\le i_1<\cdots<i_{k-1}\le N\\ i\neq i_1,\ldots,i_{k-1}}} \big(H(X_i,X_{i_1},\ldots,X_{i_{k-1}})-H(X_{i_1},\ldots,X_{i_{k-1}})\big)\\
&= \frac{k}{N\binom{N-1}{k-1}}\sum_{1\le i_1<\cdots<i_k\le N} H(X_{i_1},\ldots,X_{i_k})\\
&\quad -\frac{N-(k-1)}{N\binom{N-1}{k-1}}\sum_{1\le i_1<\cdots<i_{k-1}\le N} H(X_{i_1},\ldots,X_{i_{k-1}})\\
&= \frac{1}{\binom{N}{k}}\sum_{1\le i_1<\cdots<i_k\le N} H(X_{i_1},\ldots,X_{i_k})-\frac{1}{\binom{N}{k-1}}\sum_{1\le i_1<\cdots<i_{k-1}\le N} H(X_{i_1},\ldots,X_{i_{k-1}})\\
&= H(k)-H(k-1).
\end{aligned}$$

■

Proof. (Proposition 3.3)

$$\begin{aligned}
&\frac{1}{N(N-1)}\sum_{\substack{1\le i,j\le N\\ i\neq j}}\frac{1}{\binom{N-2}{k-1}}\sum_{\substack{1\le i_1<\cdots<i_{k-1}\le N\\ i,j\neq i_1,\ldots,i_{k-1}}} I(X_i;X_j|X_{i_1},\ldots,X_{i_{k-1}})\\
=\ &\frac{1}{N(N-1)}\sum_{\substack{1\le i,j\le N\\ i\neq j}}\frac{1}{\binom{N-2}{k-1}}\sum_{\substack{1\le i_1<\cdots<i_{k-1}\le N\\ i,j\neq i_1,\ldots,i_{k-1}}} \big(H(X_i|X_{i_1},\ldots,X_{i_{k-1}})-H(X_i|X_j,X_{i_1},\ldots,X_{i_{k-1}})\big)\\
=\ &\frac{1}{N}\sum_i\frac{N-k}{(N-1)\binom{N-2}{k-1}}\sum_{\substack{1\le i_1<\cdots<i_{k-1}\le N\\ i\neq i_1,\ldots,i_{k-1}}} H(X_i|X_{i_1},\ldots,X_{i_{k-1}})\\
&-\frac{1}{N}\sum_i\frac{k}{(N-1)\binom{N-2}{k-1}}\sum_{\substack{1\le i_1<\cdots<i_k\le N\\ i\neq i_1,\ldots,i_k}} H(X_i|X_{i_1},\ldots,X_{i_k})\\
=\ &h(k)-h(k+1).
\end{aligned}$$

■

Proof. (Lemma 3.4)

$$H(k) \;=\; \sum_{i=1}^{k} h(i) \ge k\, h(k) \ge k\, h(k+1).$$

■

Proof. (Corollary 3.5)

$$\begin{aligned} C^{(k)} - C^{(k-1)} &= \frac{N}{k}H(k) - H(N) - \frac{N}{k-1}H(k-1) + H(N) \\ &= \frac{N}{k}\left(H(k) - H(k-1) - \frac{1}{k-1}H(k-1)\right) \\ &= \frac{N}{k}(h(k) - \frac{1}{k-1}H(k-1)) \\ &\leq 0 \quad \text{since } H(k-1) \geq (k-1)\, h(k). \end{aligned}$$

■

Proof. (Proposition 3.6)
The distribution in (23) factorizes according to the set $\mathfrak{A}_{N,k+1}$. Therefore, according to Lemma 5.1 we have to prove that the A-marginal of the distribution in (23) coincides with the A-marginal of p for all $A \in \mathfrak{A}_{N,k+1}$. Let $s \geq k+1$ and $r = s-k$, that is $s - r = k$.

$$\begin{aligned} & \sum_{x_1,\ldots,x_{r-1}} \sum_{x_{s+1},\ldots,x_N} p(x_1,\ldots,x_{k+1}) \prod_{i=2}^{N-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \\ = & \sum_{x_1,\ldots,x_{r-1}} p(x_1,\ldots,x_{k+1}) \prod_{i=2}^{s-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \underbrace{\sum_{x_{s+1},\ldots,x_N} \prod_{i=s-k+1}^{N-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1})}_{=1} \\ = & \sum_{x_2,\ldots,x_{r-1}} \left(\sum_{x_1} p(x_1,\ldots,x_{k+1})\right) \prod_{i=2}^{s-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \\ = & \sum_{x_2,\ldots,x_{r-1}} p(x_2,\ldots,x_{k+1}) \prod_{i=2}^{s-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \\ = & \sum_{x_2,\ldots,x_{r-1}} p(x_2,\ldots,x_{k+1})\, p(x_{k+2}|x_2,\ldots,x_{k+1}) \prod_{i=3}^{s-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \\ = & \sum_{x_3,\ldots,x_{r-1}} \left(\sum_{x_2} p(x_2,\ldots,x_{k+2})\right) \prod_{i=3}^{s-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \\ = & \sum_{x_3,\ldots,x_{r-1}} p(x_3,\ldots,x_{k+2}) \prod_{i=3}^{s-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \\ \vdots & \qquad\qquad \vdots \\ = & \left(\sum_{x_{r-1}} p(x_{r-1},\ldots,x_{k+r-1})\right) \prod_{i=r}^{s-k} p(x_{k+i}|x_i,\ldots,x_{k+i-1}) \\ = & \; p(x_r,\ldots,x_{k+r-1})\, p(x_s|x_r,\ldots,x_{k+r-1}) \\ = & \; p(x_r,\ldots,x_s). \end{aligned}$$

Equation (24) is a direct implication of (23). ■

Proof. (Proposition 3.7)

$$
\begin{aligned}
& D(p^{(k+1)} \,\|\, p^{(k)}) \\
= \;& D(p \,\|\, p^{(k)}) - D(p \,\|\, p^{(k+1)}) \\
= \;& \sum_{i=2}^{N-k} I_p(X_{[1,i]}; X_{k+i} | X_{[i+1,k+i-1]}) + I_p(X_1; X_{k+1} | X_{[2,k]}) - \sum_{i=1}^{N-k-1} I_p(X_{[1,i]}; X_{k+i+1} | X_{[i+1,k+i]}) \\
= \;& \sum_{i=1}^{N-k-1} I_p(X_{[1,i+1]}; X_{k+i+1} | X_{[i+2,k+i]}) + I_p(X_1; X_{k+1} | X_{[2,k]}) - \sum_{i=1}^{N-k-1} I_p(X_{[1,i]}; X_{k+i+1} | X_{[i+1,k+i]}) \\
= \;& \sum_{i=1}^{N-k-1} \Big\{ \Big(H_p(X_{k+i+1} | X_{[i+2,k+i]}) - H_p(X_{k+i+1} | X_{[1,k+i]}) \Big) \\
& \qquad - \Big(H_p(X_{k+i+1} | X_{[i+1,k+i]}) - H_p(X_{k+i+1} | X_{[1,k+i]}) \Big) \Big\} + I_p(X_1; X_{k+1} | X_{[2,k]}) \\
= \;& \sum_{i=1}^{N-k-1} I_p(X_{k+i+1}; X_{i+1} | X_{[i+2,k+i]}) \;\; + \;\; I_p(X_1; X_{k+1} | X_{[2,k]}) \\
= \;& \sum_{i=0}^{N-k-1} I_p(X_{k+i+1}; X_{i+1} | X_{[i+2,k+i]}) \\
= \;& \sum_{i=1}^{N-k} I_p(X_{k+i}; X_i | X_{[i+1,k+i-1]}) \\
= \;& (N-k)\, I_p(X_1; X_{k+1} | X_{[2,k]}) \qquad \text{(if stationarity is assumed).}
\end{aligned}
$$

■