A Unifying Framework for Complexity Measures of Finite Systems

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

{nay, bertschi, jjost, 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 unifying approach for complexity measures, 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. 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.

Contents

Date: August 9, 2006.

1. Introduction

Although the paradigm of complexity turned out to be very fruitful in qualitatively understanding emergent phenomena in physical, biological, and social systems, the diversity of corresponding formalizations makes it difficult to approach a unified theory of complexity. On the other hand, in order to give complex systems research an adequate methodology that is both general and flexible, it becomes ever more important to work towards such a theory, and, in particular, to quantify complexity. We believe that new ways of looking at the whole spectrum of approaches could allow for its natural integration into a general scheme. In this article we provide some support for this belief by applying ideas from information geometry [Am, Ay1, AW, AK, MA], in order to discuss several complexity measures for finite random fields from a unifying perspective.

Many researchers agree that a system is complex if it exhibits different structures at different levels of description or if a more detailed description reveals new structural properties. It is not unambiguously clear, however, in what sense structures should be compared and which structures then should count as complex, or more complex than others, and consequently, many different notions of complexity have been proposed. Nevertheless, there do exist some insightful and quantitative approaches in certain fields on which one may hope to build. For time series, for instance, measures have been developed that capture the descriptional complexity of the underlying process [Gr, CF, BNT]. But also inspiration from properties of biological and cognitive systems has lead to formal treatments of complexity, e.g. in [TSE1].

We now discuss the organization of our paper and its main results: In Section 2 we review some information-theoretic quantities that are based on the fundamental concept of entropy and formalized in terms of stochastic processes and therefore allow for an information theoretic interpretation. In Section 3 we use them to discuss some important and well-known measures from complex-systems theory in a unifying way within the setting of finite random fields. This unified view leads to some new relations. In particular, the complexity measure by Tononi, Sporns and Edelman [TSE1] turns out to be related to the more established notions of multi-information and excess entropy. The Sections 4 and 5 represent the second part of the paper, where we develop a geometric interpretation for complexity in the spirit of these complexity measures from Section 3 within the setting of information geometry. This leads to a general geometric structure for complexity measures that is based on a clear distinction of the interaction orders of the nodes in the system. The geometric framework thus expresses the idea that complexity requires interactions at different scales of description by using an explicit decomposition in terms of an interaction hierarchy.
We can thus quantify the insight that *complex systems are more than the sum of their parts, and not just more than the sum of their elements*.

2. Preliminaries from Information Theory

In this section, we recall some basic notions from information theory; a reference is [CT]. We consider a set V of $1 \leq N < \infty$ nodes with state sets $\mathcal{X}_v$, $v \in V$. Given a finite subset $A \subseteq V$, we write $\mathcal{X}_A$ instead of $\times_{v\in A}\mathcal{X}_v$, and the total configuration set is $\mathcal{X}_V$. $|\mathcal{X}_A|$ is the number of elements in $\mathcal{X}_A$, that is, the number of different states that can be attained on A. We have the natural projections

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

Given a probability vector p on $\mathcal{X}_V$, the X_A become random variables. For three subsets A, B, C of V, we shall use the following information-theoretic quantities: The *entropy* of X_C is defined as

$$H_p(X_C) \;:=\; -\sum_{z\in\mathcal{X}_C} \Pr(X_C = z)\log_2\big(\Pr(X_C = z)\big)$$

This quantity is a natural measure of the uncertainty that one has about the outcome of X_C, that is, the information one expects to gain by observing that outcome. The maximal value for the entropy is $\log_2(|\mathcal{X}_V|)$. If we subtract the entropy from this maximal value, we get a measure for the information about the outcome of X_C contained in the probability vector p: $\log_2(|\mathcal{X}_V|) - H_p(X_C)$. Knowing the outcome of X_B reduces the uncertainty that one has about the outcome of X_C. The remaining uncertainty is then quantified by the *conditional entropy* of X_C given X_B:

$$H_p(X_C \,|\, X_B) \;:=\; -\sum_{y\in\mathcal{X}_B,\, z\in\mathcal{X}_C} \Pr(X_B = y,\, X_C = z)\log_2\big(\Pr(X_C = z \,|\, X_B = y)\big)$$

In terms of these entropy measures, the *mutual information* of X_C and X_B is given by

$$I_p(X_C : X_B) \;:=\; H_p(X_C) - H_p(X_C \,|\, X_B)$$

which measures the reduction of the uncertainty of the outcome of X_C given the outcome of X_B and vice versa.
The *conditional mutual information*, which is defined as

$$I_p(X_C : X_B|X_A) \;:=\; H_p(X_C|X_A) - H_p(X_C|X_A, X_B)\;,$$

quantifies the reduction of the uncertainty of the outcome of X_C given the outcome of X_B, if the outcome from X_A was already known.
Having a quantity that depends on p, say $H_p(X_V)$, we denote the corresponding function by the same symbol without specifying p. For example, we write $H(X_V)$ for the function $p \mapsto H_p(X_V)$.

3. Information-Theoretic Complexity Measures

The information-theoretic quantities of Section 2 can be used to define some complexity measures for random fields: We shall define multi-information and excess-entropy as such measures and embed them into the more general setting of TSE complexity.

3.1. **Multi-information.** The multi-information is defined as

$$I_p(X_V) = \sum_{v\in V} H_p(X_v) - H_p(X_V) \tag{1}$$

$$= \Big(\log_2(|\mathcal{X}_V|) - H_p(X_V)\Big) - \sum_{v\in V}\Big(\log_2(|\mathcal{X}_v|) - H_p(X_v)\Big) \tag{2}$$

It quantifies the total statistical interdependence of the nodes with respect to the joint distribution p. In the literature it is also referred to as *redundancy*, *integration*, and *complexity* [St, SV, TSE1, Mo, AW]. It becomes zero if and only if the probability distribution p has the product structure

$$p(x) = \prod_{v\in V} p_v(x_v)\ , \tag{3}$$

where each p_v denotes the image distribution of the projection X_v. In particular, the multi-information vanishes in the case of complete randomness, given by the uniform distribution, and in the case of complete determinism, given by a distribution that is concentrated in one configuration. This property is frequently stated as a necessary but not sufficient requirement for a quantity to be considered as a complexity measure. Moreover, the representation (2) shows that the multi-information quantifies the excess of the system's total information in relation to the sum of its elements' informations. The difference structure of this quantity is perfectly consistent with the concept that *complex systems are more than the sum of their elements*. This provides further support for considering the multi-information as a natural candidate for a complexity measure. On the other hand, maximizing the multi-information leads to probability distributions that are in some sense very simple and can hardly be considered as complex. The general structure of maximizers of the multi-information has been studied in [AK]. To simplify the presentation, we consider systems with elements $v \in \{1,\dots,N\} = V$ and corresponding state sets $\mathcal{X}_v = \{0,1\}$ for all $v \in V$. A global maximizer of the multi-information is then given by a family $\sigma_i : \{0,1\} \to \{0,1\}$, $i = 2,\dots,N$, of bijective maps. They allow for an embedding $\sigma : \{0,1\} \to \{0,1\}^N = \mathcal{X}_V$, $x \mapsto (x, \sigma_2(x), \dots, \sigma_N(x))$. The image $\mathcal{C}_\sigma$ of this map consists of exactly two elements c_1 and c_2, and the the corresponding uniform distribution on this code is given by $\frac{1}{2}(\delta_{c_1} + \delta_{c_2})$. It is not hard to see that the global maximizers of the multi-informations are exactly these 2^{N-1} distributions [AK].

Any sequential decomposition of the system is consistent with a corresponding decomposition of the multi-information into a sum of mutual informations. In order to make this point more precise, we describe a natural way of constructing a sequence of partitions. The idea is to divide the system into finer and finer partitions according to the following rule:

(1) *Initialization:* Start the sequence of partitions by defining as first partition the trivial one: $\xi_1 := \{V\}$
(2) *Step $k \to k+1$:* If all atoms of the partition ξ_k have exactly one element, then stop. Otherwise, choose one atom A_k of the partition ξ_k that has at least two elements and divide it into two non-empty and disjoint sets

A_k^1 and A_k^2 with $A_k = A_k^1 \cup A_k^2$. Define the new partition ξ_{k+1} according to

$$\xi_{k+1} \;:=\; \big(\xi_k \setminus \{A_k\}\big) \cup \{A_k^1, A_k^2\}$$

(3) Go to the second step.

This procedure generates a sequence of bipartitions $A_k = A_k^1 \cup A_k^2$, and for all k we have the decomposition rule

$$I(X_{A_k}) \;=\; I(X_{A_k^1} : X_{A_k^2}) \,+\, I(X_{A_k^1}) \,+\, I(X_{A_k^2})\;, \tag{4}$$

which finally leads to the chain rule for multi-information

$$I(X_V) \;=\; \sum_{k=1}^{N-1} I(X_{A_k^1} : X_{A_k^2})\;. \tag{5}$$

In the next section we will introduce a second dependence measure that has similar decomposition properties.

3.2. **Excess entropy.** The excess entropy was originally introduced under the name *effective measure complexity* [Gr] in the context of time series as the minimal amount of memory required for an optimal prediction. Closely related measures are the statistical complexity of ϵ-machines proposed in [CY] and the predictive information [BNT]. In a recent overview [CF], the effective measure complexity was termed *excess entropy* which we will use in the following because it relates directly to one definition of this quantity.
In a time series, 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 $(x_1, x_2, \ldots) \mapsto (x_2, x_3, \ldots)$. This allows for identifying limits of sequences in the way we are going to do.
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,

$$h_\infty \;:=\; \lim_{N\to\infty} h_N \tag{6}$$

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

$$E \;:=\; \lim_{N\to\infty} \left(H(X_1, \ldots, X_N) - N h_\infty\right) \tag{7}$$

It measures the nonextensive part of the entropy, i.e. the amount of entropy of each element that *exceeds* the entropy rate.
The excess entropy of a finite string can be defined as

$$E_N \;:=\; H(X_1, \ldots, X_N) - N h_N \tag{8}$$

$$\;=\; \sum_{k=1}^{N} (h_k - h_N) \tag{9}$$

One straightforward generalization to a finite system V can be made by assuming an arbitrary order $v_1, v_2, \ldots, v_N$ of the elements leading to

$$E(X_V) = \sum_{k=1}^{N} \left(H(X_{v_k}|X_{v_{k-1}}, \ldots, X_{v_1}) - H(X_{v_k}|X_{V\setminus\{v_k\}}) \right) \tag{10}$$

$$= H(X_V) - \sum_{v\in V} H(X_v|X_{V\setminus\{v\}}) \tag{11}$$

$$= (N-1)\left(\frac{1}{N-1} \sum_{v\in V} H(X_{V\setminus\{v\}}) \; - \; H(X_V) \right) \tag{12}$$

The representation (11) shows that the excess entropy is independent of the initially chosen order of the elements. One can also regard (11) as generalization of (8) with h_N replaced by the conditional entropy $H(X_v|X_{V\setminus\{v\}})$ averaged over all v.

The excess entropy shares some features of the multi-information. As we see from (12), it has, up to a scaling factor, an entropy difference structure similar to the structure (1) of the multi-information. The excess entropy also has the property that it vanishes in the case of independent nodes, which have a joint distribution of the product structure (3). Furthermore, for binary units the maximal value of the excess entropy coincides with the maximal value of the multi-information, which is $N-1$. But the structure of the maximizers is characterized by the idea of a parity code instead of the idea of a repetition code, which we used for describing the maximizers of the multi-information. To be more precise, consider the set $\{0,1\}$ as a finite field, and define the following linear map between finite vector spaces: $\sigma : \{0,1\}^{N-1} \to \{0,1\}^N$, $(x_1, \ldots, x_{N-1}) \mapsto (x_1, \ldots, x_{N-1}, \sum_{i=1}^{N-1} x_i)$. Then it is not hard to see that the uniform distribution on the image of σ (parity code) globally maximizes the excess entropy.

In fact, within the time series context, the excess entropy has been considered as a distinguished complexity measure, determined by learning-theoretical assumptions [BNT]. For the case of finite systems this complexity measure was mentioned in passing in [TSE2]. The excess entropy is monotonically increasing with the system size because

$$E(X_{V\cup\{v_{N+1}\}}) - E(X_V) \; = \; \sum_{i=1}^{N} I(X_{v_i} : X_{v_{N+1}}|X_{V\setminus\{v_i\}}) \geq 0 \tag{13}$$

Using the notation introduced in Section 3.1, we have the following decomposition similar to (5):

$$E(X_V) \; = \; \sum_{k=1}^{N-1} I(X_{A_k^1} : X_{A_k^2}|X_{V\setminus(A_k^1\cup A_k^2)}) \, . \tag{14}$$

This can be proven by induction introducing the excess entropy $E(\xi_k)$ of the partition ξ_k as

$$E(\xi_k) := H(X_V) - \sum_{A\in\xi_k} H(X_A|X_{V\setminus A}) \, .$$

For $k = 2$ we have

$$\begin{aligned} E(\xi_2) &= H(X_{A_1^1}, X_{A_1^2}) - H(X_{A_1^1}|X_{A_1^2}) - H(X_{A_1^2}|X_{A_1^1}) \\ &= I(X_{A_1^1} : X_{A_1^2}) \,. \end{aligned}$$

Now let us consider the step from partition ξ_k to a partition ξ_{k+1}. The excess entropy with respect to the finer partition ξ_{k+1} is given as

$$\begin{aligned} & E(\xi_{k+1}) \\ = \;& H(X_V) - \sum_{A\in\xi_{k+1}} H_p(X_A|X_{V\setminus A}) \\ = \;& H(X_V) - \sum_{A\in\xi_k\setminus\{A_k\}} H(X_A|X_{V\setminus A}) - H(X_{A_k^1}|X_{V\setminus A_k^1}) - H(X_{A_k^2}|X_{V\setminus A_k^2}) \\ = \;& H(X_V) - \sum_{A\in\xi_k} H(X_A|X_{V\setminus A}) \\ & \qquad - H(X_{A_k^1}|X_{V\setminus A_k^1}) - H(X_{A_k^2}|X_{V\setminus A_k^2}) + H(X_{A_k^1\cup A_k^2}|X_{V\setminus A_k}) \\ = \;& E(\xi_k) + I(X_{A_k^1} : X_{A_k^2}|X_{V\setminus\{A_k^1\cup A_k^2\}}) \,. \end{aligned}$$

This proves the chain rule (14). In comparison to the chain rule for the multi-information the mutual informations are now conditioned on the rest of the system, which means that only the new information at each step of refining the partition is considered. Note that nevertheless it is possible that $E_p(X_V) > I_p(X_V)$.

3.3. **TSE complexity.** The multi-information and the excess entropy can be considered as extreme cases of a general definition that appears in a complexity measure introduced by Tononi, Sporns, and Edelman [TSE1]. In order to define the TSE complexity, we first introduce

$$C^{(k)}(X_V) := I(X_V) - \frac{N}{k\binom{N}{k}} \sum_{\substack{A\subseteq V\\|A|=k}} I(X_A) \tag{15}$$

$$= \frac{N}{k\binom{N}{k}} \sum_{\substack{A\subseteq V\\|A|=k}} H(X_A) - H(X_V) \tag{16}$$

For $k = 1$ we recover the multi-information, and for $k = N - 1$ we get, up to a constant factor, the excess entropy:

$$I(X_V) = C^{(1)}(X_V), \qquad E(X_V) = (N-1)\, C^{(N-1)}(X_V) \tag{17}$$

Furthermore, if we assume independence of the nodes, then we get

$$\begin{aligned} C_p^{(k)}(X_V) &= \frac{N}{k\binom{N}{k}} \sum_{\substack{A\subseteq V\\|A|=k}} \sum_{v\in A} H_p(X_v) - \sum_{v\in V} H_p(X_v) \\ &= \sum_{v\in V} \frac{N}{k\binom{N}{k}} \binom{N-1}{k-1} H_p(X_v) - \sum_{v\in V} H_p(X_v) \\ &= 0 \,. \end{aligned}$$

Thus, extending the two natural examples given by $k = 1$ and $k = N - 1$ to all the intermediate k's does not affect the property of vanishing in the independence case. In particular, the $C_p^{(k)}(X_V)$ vanish in the two extreme cases of complete randomness, $H_p(X_V) = \log_2(|\mathcal{X}_V|)$, and complete determinism, $H_p(X_V) = 0$. The TSE complexity is defined as a weighted sum of the $C^{(k)}(X_V)$'s:

$$C(X_V) := \sum_{k=1}^{N-1} \frac{k}{N} C^{(k)}(X_V) \tag{18}$$

Although our main focus will be on the multi-information and the excess entropy as contributions to the TSE complexity, there is another interesting relation between the TSE complexity and the excess entropy that is provided by the transient information. We address this connection in the next section.

3.4. **Transient information.** In [CF, FC] another complexity measure, the transient information, was introduced to measure the complexity of periodic sequences. The excess entropy (7) for a periodic sequence of period P is equal to $\log_2 P$. Therefore the excess entropy cannot be used to distinguish the complexity of different sequences with the same period. Using the excess entropy (7), the transient information of an infinite sequence was defined as

$$T = \sum_{k=1}^{\infty} (E + kh_\infty - H(X_1, \ldots, X_k)) \ . \tag{19}$$

Whereas the excess entropy was a bound for the amount of memory necessary for a model that provides optimal predictions the transient information quantifies the amount of information that has to be extracted from the process in order to identify the state of the system [CF].
Using the excess entropy (11) for an arbitrary system consisting of N elements, one can easily define a corresponding transient information as

$$T(X_V) = \sum_{k=1}^{N} \left(E(X_V) + \frac{k}{N} \sum_{v \in V} H(X_v | X_{V \setminus \{v\}}) - \frac{1}{\binom{N}{k}} \sum_{\substack{A \subseteq V \\ |A|=k}} H(X_A) \right) \ . \tag{20}$$

With (11) we derive the equivalent expression

$$T(X_V) = \frac{N-1}{2} E(X_V) + \sum_{k=1}^{N} \left(\frac{k}{N} H(X_V) - \frac{1}{\binom{N}{k}} \sum_{\substack{A \subseteq V \\ |A|=k}} H(X_A) \right) \ . \tag{21}$$

From (21), we see a relationship between the excess entropy and the above complexity:

$$T(X_V) + C(X_V) = \frac{N-1}{2} E(X_V) \ . \tag{22}$$

As both, $T(X_V)$ and $C(X_V)$ are non-negative quantities, we see that they can be only positive if the excess entropy $E(X_V)$ is positive. The following figure illustrates the equality (22).

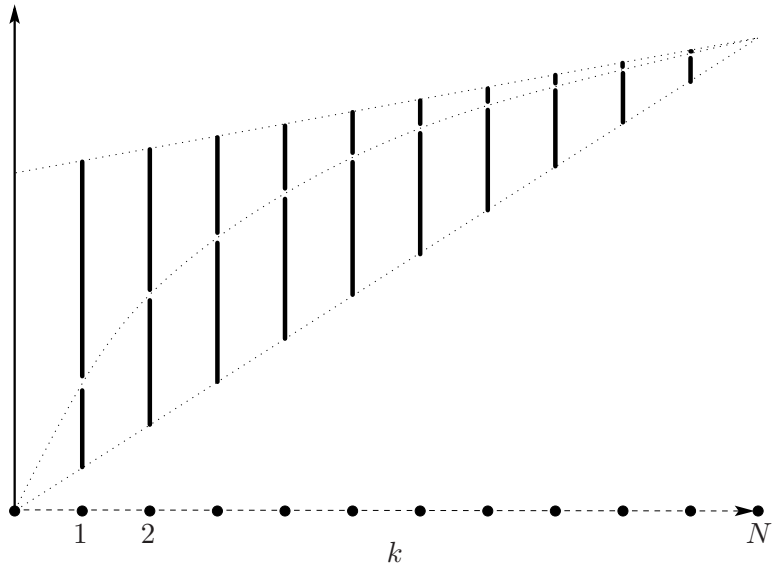

Figure 1: The curved dotted line represents the graph of the function $k \mapsto \langle H \rangle_k := \frac{1}{\binom{N}{k}} \sum_{\substack{A \subseteq V \\ |A|=k}} H_p(X_A)$. The lower dotted linear graph is given by $k \mapsto \frac{k}{N} H_p(X_V)$, and the total length of the vertical lines between these two functions is nothing but the complexity $C_p(X_V)$. The upper linear graph is given by $k \mapsto E_p(X_V) + \frac{k}{N} \sum_{v \in V} H_p(X_v | X_{V \setminus \{v\}})$, and the total length of the vertical lines between this graph and the curved graph is nothing but the transient information $T_p(X_V)$. The sum of $T_p(X_V)$ and $C_p(X_V)$ is then equal to the total length of all vertical lines between the two linear graphs, which is $\frac{N-1}{2} E(X_V)$. This is exactly the statement of (22).

4. Preliminaries from Information Geometry

4.1. **Projections onto exponential families.** Given a nonempty finite set $\mathcal{X}$, we denote the set of probability distributions on $\mathcal{X}$ by $\bar{\mathcal{P}}(\mathcal{X})$. The *support* of $p \in \bar{\mathcal{P}}(\mathcal{X})$ is defined as $\mathrm{supp}(p) := \{x \in \mathcal{X} \ : \ p(x) > 0\}$. To simplify the arguments we shall mainly consider the set $\mathcal{P}(\mathcal{X})$ of probability distributions with total support. With the exponential map

$$\exp : \ \mathbb{R}^{\mathcal{X}} \to \mathcal{P}(\mathcal{X}), \qquad f \mapsto \frac{\exp(f)}{\sum_{x \in \mathcal{X}} \exp(f(x))}$$

an *exponential family* is defined as the image $\exp(\mathcal{V})$ of a linear subspace $\mathcal{V}$ of $\mathbb{R}^{\mathcal{X}}$. The "distance" (*relative entropy*, *KL-divergence*) of two distributions p and q is measured by

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

Given an exponential family $\mathcal{E}$, the function

$$\bar{\mathcal{P}}(\mathcal{X}) \;\to\; \mathbb{R}, \qquad p \;\mapsto\; D(p\,\|\,\mathcal{E}) := \inf_{q\in\mathcal{E}} D(p\,\|\,q)$$

is continuous. For a positive distribution $p \in \mathcal{P}(\mathcal{X})$, there exists a unique distribution $\Pi_{\mathcal{E}}(p) \in \mathcal{E}$ with

$$D(p\,\|\,\Pi_{\mathcal{E}}(p)) \;=\; D(p\,\|\,\mathcal{E})$$

4.2. **Interaction spaces.** We now use the compositional structure of $\mathcal{X}_V$ in order to define exponential families given by interaction spaces. We decompose $x \in \mathcal{X}_V$ in the form $x = (x_A, x_{V\setminus A})$ with $x_A \in \mathcal{X}_A$, $x_{V\setminus A} \in \mathcal{X}_{V\setminus A}$, and define $\mathcal{I}_A$ to be the subspace of functions that do not depend on the configurations $x_{V\setminus A}$:

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

In the following we use these interaction spaces as building blocks for more general interaction spaces and associated exponential families. The most general construction is based on a set of subsets of V, a so-called *hypergraph*. Given such a set $\mathscr{A} \subseteq 2^V$, we define the corresponding interaction space by

$$\mathcal{I}_{\mathscr{A}} \;:=\; \sum_{A\in\mathscr{A}} \mathcal{I}_A$$

and consider the corresponding exponential family $\exp(\mathscr{A}) := \exp(\mathcal{I}_{\mathscr{A}})$. If two sets $\mathscr{A}_1, \mathscr{A}_2 \subseteq 2^V$ have the same maximal subsets of V as elements, then obviously $\exp(\mathscr{A}_1) = \exp(\mathscr{A}_2)$. We define the following order relation: $\mathscr{A}_1 \preceq \mathscr{A}_2$ if for all $A_1 \in \mathscr{A}_1$ there exists some $A_2 \in \mathscr{A}_2$ with $A_1 \subseteq A_2$. This implies

$$\mathscr{A}_1 \preceq \mathscr{A}_2 \qquad \Rightarrow \qquad \exp(\mathscr{A}_1) \subseteq \exp(\mathscr{A}_2) \tag{23}$$

In general, there is no explicit formula for the projections onto these exponential families and the corresponding distances. But in special situations, this is possible, and, in these geometric definitions, we recover the well known information-theoretic quantities from Section 2:

Examples.

(1) If $\mathscr{A} = \{\emptyset\}$, then

$$\Pi_{\exp(\mathscr{A})}(p)(x) \;=\; \frac{1}{|\mathcal{X}_V|}, \qquad D(p\,\|\,\exp(\mathscr{A})) \;=\; \log_2(|\mathcal{X}_V|) - H_p(X_V)$$

(2) If $\mathscr{A} = \{A\}$ with a non-empty subset A of V, then

$$\Pi_{\exp(\mathscr{A})}(p)(x) \;=\; \frac{1}{|\mathcal{X}_{V\setminus A}|}\, p(x_A)$$

$$D(p\,\|\,\exp(\mathscr{A})) \;=\; \log_2(|\mathcal{X}_{V\setminus A}|) - H_p(X_{V\setminus A}\,|\,X_A)$$

(3) Let $\{A, B, C\}$ be a partition of V into three subsets, and $\mathscr{A} := \{A \cup C, B \cup C\}$. Then

$$\Pi_{\exp(\mathscr{A})}(p)(x) \;=\; \frac{p(x_A, x_C)\, p(x_B, x_C)}{p(x_C)}$$

$$D(p \,\|\, \exp(\mathscr{A})) \;=\; I_p(X_A : X_B \,|\, X_C)$$

(4) Let $\mathscr{A} = \{A_1, \ldots, A_r\}$ be a partition of V. Then

$$\Pi_{\exp(\mathscr{A})}(p)(x) \;=\; \prod_{i=1}^{r} p(x_{A_i})$$

$$D(p \,\|\, \exp(\mathscr{A})) \;=\; \sum_{i=1}^{r} H_p(X_{A_i}) \;-\; H_p(X_V)$$

As a special case for $r = 2$ we recover the mutual information.

4.3. **Hierarchies of exponential families.** For an increasing sequence

$$\mathscr{A}_1 \preceq \mathscr{A}_2 \preceq \cdots \preceq \mathscr{A}_L$$

of sets $\mathcal{A}_k \subseteq 2^V$, $k = 1, \ldots, L$, by (23) one has an associated hierarchy of exponential families:

$$\exp(\mathscr{A}_1) \subseteq \exp(\mathscr{A}_2) \subseteq \cdots \subseteq \exp(\mathscr{A}_L).$$

With $p^{(k)}$ we denote the projection of p onto the exponential family $\exp(\mathscr{A}_k)$. Then the Pythagoras theorem for the relative entropy implies the following decomposition:

$$D(p \,\|\, p^{(1)}) \;=\; \sum_{k=2}^{L} D(p^{(k)} \,\|\, p^{(k-1)}) \;+\; D(p \,\|\, p^{(L)}). \tag{24}$$

Applied to specific hierarchies this leads to some well-known chain rules.

Examples (Chain Rules).

(1) We start with the following example $V := \{1, \ldots, N\}$,

$$A_k := \{1, \ldots, k-1\}, \;\; \mathscr{A}_k := \{A_k\}, \quad k = 1, 2, 3, \ldots, N.$$

In this case the decomposition (24) becomes

$$\log_2(|\mathcal{X}_V|) - H_p(X_V) \;=\; \sum_{k=1}^{N} \big(\log_2(|\mathcal{X}_k|) \;-\; H_p(X_k | X_1, \ldots, X_{k-1}) \big).$$

This is equivalent to the chain rule for the entropy

$$H_p(X_V) \;=\; \sum_{k=1}^{N} H_p(X_k | X_1, \ldots, X_{k-1}).$$

(2) Consider the sequence ξ_k, $k = 1, \ldots, N$, of partitions that we defined in Section 3.1. Then (24) becomes the chain rule (5) for the multi-information.

5. The Geometric View of Complexity

Here we consider a hierarchy of exponential families that has been studied by Amari [Am].

$$\mathscr{A}_k := \{A \subseteq V \,:\, |A| = k\}, \qquad k = 0, \ldots, N.$$

This implies the following hierarchy of exponential families:

$$\mathcal{F}^{(k)} \;:=\; \exp(\mathscr{A}_k), \qquad k = 0, \ldots, N.$$

Here, $\mathcal{F}^{(0)}$ contains exactly one element, namely the center of the simplex (uniform distribution), $\mathcal{F}^{(1)}$ is the set of factorized distributions (complete independence), and $\mathcal{F}^{(N)}$ coincides with the whole simplex $\mathcal{P}(\mathcal{X}_V)$. The distance $D(p \,\|\, p^{(1)})$ is nothing but the multi-information among the units, which we have already introduced in Section 3.1. As we have seen, the maximizers of the multi-information have a very simple structure. It turns out that they are all contained in the topological closure of $\mathcal{F}^{(2)}$ [AK]. This fact indicates a geometric reason why multi-information should not be considered as the right complexity measure. In order to make this point clear we have to decompose the multi-information into the contributions $D(p^{(k)} \,\|\, p^{(k-1)})$, $k = 2, \ldots, N$, to the stochastic interdependence that cannot be explained by interactions of order $\leq k-1$ [Am]. More precisely, we have

$$D(p \,\|\, p^{(1)}) \;=\; \sum_{k=2}^{N} D(p^{(k)} \,\|\, p^{(k-1)}).$$

The fact that all global maximizers of the multi-information are contained in the closure of $\mathcal{F}^{(2)}$ implies that they are completely characterized by second-order marginals, and therefore we have $D(p^{(k)} \,\|\, p^{(k-1)}) = 0$ for all $3 \leq k \leq N$ in that case. This means that the maximization of the sum of all the contributions $D(p^{(k)} \,\|\, p^{(k-1)})$, which is the multi-information, leads to just one contribution, namely $D(p^{(2)} \| p^{(1)})$, which has the large value $(N-1)\log_2(2)$. If we interpret $p^{(2)}$ as a "sum" of all marginals p_A, $|A| \leq 2$, then p is nothing but the sum of its parts of size ≤ 2. In order to quantify complexity, this observation suggests to relate the whole not just to its elements (parts of size one) but to all its constituent parts of any size. In this sense, *complex systems are more than the sum of their parts, and not just more than the sum of their elements*. Therefore, we interpret $D(p \,\|\, p^{(k)})$ as the distance of p from the sum of its parts p_A, where $A \subseteq V$ has size $\leq k$, and consider a weighted sum of these distances as a general structure for complexity measures. More precisely, with a weight vector $\alpha = (\alpha(1), \ldots, \alpha(N-1)) \in \mathbb{R}^{(N-1)}$ we set:

$$C_\alpha(p) \;:=\; \sum_{k=1}^{N-1} \alpha(k)\, D(p \,\|\, p^{(k)}) \tag{25}$$

$$\;=\; \sum_{k=2}^{N} \left(\sum_{i=1}^{k-1} \alpha(i) \right) D(p^{(k)} \,\|\, p^{(k-1)}) \tag{26}$$

As we have seen, the multi-information can be represented in this way by setting $\alpha(1) := 1$, and $\alpha(k) := 0$ for $k \geq 2$. 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 $D(p^{(k)} \,\|\, p^{(k-1)})$ to display a specific shape of behaviour as k increases. A similar intuition is reflected by the structure (18) of the TSE complexity, which we have discussed in Section 3.3. Within the geometric framework, this comparison suggests to relate the $C_p^{(k)}(X_V)$ to our distances $D(p \,\|\, p^{(k)})$, which appear in the sum (25). Interpreting the representation (15) as deviation of the total stochastic dependence $I_p(X_V)$ from the dependencies up to order k one would expect that there is a close relation between the two quantities $C_p^{(k)}(X_V)$ and $D(p \,\|\, p^{(k)})$. In order to have a better understanding of their relation, we consider the special case where N is a multiple of k, say $N = r \cdot k$ with $r \in \mathbb{N}$, and a corresponding partition $\xi = \{A_1, \ldots, A_r\}$ of V into r subsets with cardinality k (*k-partition*). For ξ we obviously have $D(p \,\|\, \mathcal{F}^{(k)}) \leq D(p \,\|\, \exp(\xi))$. Denoting the total number $\frac{1}{r!}\binom{N}{k}\binom{N-k}{k}\cdots\binom{N-(r-1)k}{k}$ of k-partitions by L, this directly implies

$$D(p \,\|\, \exp\Big(\bigcup \xi\Big)) \;=\; D(p \,\|\, \mathcal{F}^{(k)}) \;\leq\; \frac{1}{L}\sum D(p \,\|\, \exp(\xi)) \;. \tag{27}$$

Here, the union on the left-hand side and the sum on the right-hand side are taken over all k-partitions ξ. It turns out that the mean on the right-hand side is nothing but $C_p^{(k)}(X_V)$:

$$\begin{aligned}
& \frac{1}{L} \sum_{\substack{\xi \\ k\text{-partition}}} D(p \,\|\, \exp(\xi)) \\
= \;& \frac{1}{L} \sum_{\substack{\xi=\{A_1,\ldots,A_r\} \\ k\text{-partition}}} I_p(X_{A_1}, \ldots, X_{A_r}) \\
= \;& \frac{1}{L} \sum_{\substack{\xi=\{A_1,\ldots,A_r\} \\ k\text{-partition}}} \sum_{i=1}^{r} H_p(X_{A_i}) \;-\; H_p(X_V) \\
= \;& \frac{1}{L} \sum_{\substack{A\subseteq V \\ |A|=k}} \frac{r}{r!} \binom{N-k}{k} \cdots \binom{N-(r-1)k}{k} H_p(X_A) \;-\; H_p(X_V) \\
= \;& \frac{N}{k\binom{N}{k}} \sum_{\substack{A\subseteq V \\ |A|=k}} H_p(X_A) \;-\; H_p(X_V) \\
= \;& C_p^{(k)}(X_V)
\end{aligned}$$

The inequality (27) can now be written as

$$D(p \,\|\, p^{(k)}) \;\leq\; C_p^{(k)}(X_V) \;, \tag{28}$$

and it represents a convexity-like property, which provides some intuition on the difference between the TSE complexity and the following geometric modification

of it:

$$\begin{aligned} C_p^*(X_V) &:= \sum_{k=1}^{N-1} \frac{k}{N} D(p \,\|\, p^{(k)}) \\ &= \sum_{k=2}^{N} \frac{(k-1)k}{2N} D(p^{(k)} \,\|\, p^{(k-1)}) \end{aligned}$$

This modification corresponds to (25) with $\alpha(k) = \frac{k}{N}$.

6. 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 by marginal entropies, because those usually do not provide a complete decomposition of stochastic dependence. 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. While well-known complexity measures such as multi-information, excess entropy, or the TSE complexity are defined in terms of marginal entropies we propose a family of complexity measures (25) that is directly linked to this "orthogonal" decomposition of the stochastic dependence. Although we think that the corresponding hierarchy of exponential families plays a distinguished role within the analysis of complexity, the general information-geometric method allows for studying other hierarchies and offers both flexibility of application and a comprehensive view of complexity. We hope that this geometric view will be used to identify and analyze new "dimensions" of complexity in a transparent manner.

Acknowledgements

All authors are grateful to Sidney Frankel for a fruitful collaboration within the systems theory group at the MPI for Mathematics in the Sciences. Nihat Ay thanks Andreas Knauf, David Krakauer, and Eric Smith for discussions on the subject of complexity from the point of view of mathematical physics and biololgy.

References

[Am] S. Amari. *Information geometry on hierarchy of probability distributions*, IEEE Trans. IT **47**, 1701-1711 (2001)

[Ay1] N. Ay. *An information-geometric approach to a theory of pragmatic structuring*, Ann. Prob. **30**, 416-436 (2002)

[AW] N. Ay, T. Wennekers. *Dynamical Properties of Strongly Interacting Markov Chains*. Neural Networks 16, 1483–1497 (2003).

[AK] N. Ay, A. Knauf. *Maximizing multi-information*, Kybernetika (2006), in press

[BNT] W. Bialek, I. Nemenman, N. Tishby. *Predictability, Complexity, and Learning*, Neural Computation 13, 2409-2463 (2001)

[CT] T. Cover, J. Thomas, *Elements of Information Theory*, Wiley, 1991

[CY] J. P. Crutchfield and K. Young. *Inferring Statistical Complexity*, Phys. Rev. Lett. 63, 105-108 (1989)

[CF] J. P. Crutchfield and David P. Feldman, *Regularities unseen, randomness observed: Levels of entropy convergence*, Chaos 13 (1), 25–54 (2003)

[FC] D. P. Feldman, J. P. Crutchfield, *Synchronizing to Periodicity: The Transient information and Synchronization Time of Periodic Sequences*, Adv Compl Sys 7, 329–355 (2004)

[Gr] P. Grassberger. *Toward a quantitative theory of self-generated complexity*, Int. J. Theor. Phys. 25 (9), 907–938 (1986)

[Mo] S. D. Morgera. *Information Theoretic Covariance Complexity and its Relation to Pattern Recognition*, IEEE Transactions on Systems, Man, and Cybernetics 15 (5), 608–619 (1985)

[MA] F. Matus, N. Ay. *On maximization of the information divergence from an exponential family*, Proceedings of WUPES03 (ed. J. Vejnarova), University of Economics Prague, 199-204 (2003)

[St] M. Studeny. *Probabilistic Conditional Independence Structures*. Series: Information Science and Statistics, Springer 2005.

[SV] M. Studeny, J. Vejnarova. *The multiinformation function as a tool for measuring stochastic dependence*. In M. I. Jordan (ed.) 1998. *Learning in Graphical Models*, Dordrecht: Kluwer, 1998.

[TSE1] 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**, 5033-5037 (1994)

[TSE2] G. Tononi, O. Sporns, G. M. Edelman, *A measure for brain complexity: Relating functional segregation and integration in the nervous system*, PNAS 96, 3257–3267 (1999)