Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Information and Complexity, or: Where is the Information?

by

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

Preprint no.: 102 2020

Information and Complexity, or: Where is the Information?

Nihat Ay[*], Nils Bertschinger[†], Jürgen Jost[‡], Eckehard Olbrich[§], Johannes Rauh[¶]

October 28, 2020

Contents

[*]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, and Santa Fe Institute for the Sciences of Complexity, New Mexico, USA
[†]Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany, and Goethe University, Frankfurt am Main, Germany
[‡]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, and Santa Fe Institute for the Sciences of Complexity, New Mexico, USA
[§]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[¶]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

1 Introduction

The concepts of information and complexity seem to be intricately linked. Complexity notions are quantified in information theoretical terms, and a general principle might say that a structure is the more complex, the more information is needed to describe or build it. That principle, however, needs some qualification. One should distinguish between – usually useful – information about regularities of a structure or a process and – often useless – information about random details. The question is not only *information about what?*, but also *where is that information?*, that is, whether and how it is or can be internally stored in a system with limited capacity, at which level of a process information is needed to predict the continuation of a process, and where it can be found in a distributed system. In the latter case, we should, however, not only look for information that is exclusively located somewhere or that is shared between entities, but should also consider complementary or synergistic information, that is, information that only emerges when several sources are combined.

These lecture notes describe what is currently known about these questions, and they develop the underlying theoretical concepts and elucidate them at simple examples. Also, when we can quantify complexity concepts, we can also try to optimize the corresponding complexity measures. This will also be systematically discussed.

These notes are the result of a series of lectures that one of us (JJ) delivered at the Summer School in Como in July, 2018. They present work that we have done jointly during the last few years. JJ thanks Elisa Mastrogiacomo and Sergio Albeverio for organizing a very stimulating school, and the participants and the other lecturers, in particular Luciano Boi, Ivar Ekeland and Frank Riedel, for stimulating discussions.

2 Background: Principles of information theory

2.1 Shannon information

The basic concept is that of the *Shannon Information* [Shannon 1948] of a random variable X, or equivalently, of a probability distribution p, when the possible values x_i of X are realized with probabilities $p_i = p(x_i)$. These probabilities satisfy $0 \leq p_i \leq 1$ for all i, with the normalization $\sum_i p_i = 1$. The Shannon information or entropy then is

$$H(X) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i \text{ (bits)}. \tag{1}$$

This is the expected *reduction of uncertainty*, i.e., the *information gain*, if we learn which concrete value x_i of the random variable X from a known distribution p with probabilities $p_i = p(x_i)$ is realized.
This is the basic example: When we have two possible events occurring with

equal probability 1/2 (an unbiased coin) we thus gain $\log_2 2 = 1$ bit of information when we observe the outcome.
A fair dice yields $\log_2 6$ bits of information.

2.2 Mutual and conditional information

We now consider the situation where we have an additional random variable Y. In the example of the dice, we could let $Y = 0$ (resp. 1) for an odd (even) result, each with probability 1/2. According to the basic example, we have $H(Y) = 1$ bit. When we know Y, there remain only 3 possibilities for the value of X, each with probability 1/3.
This leads us to the concept of *conditional information*; in this example, the remaining uncertainty about X when knowing Y is

$$H(X|Y) = \log_2 3. \tag{2}$$

Thus, the uncertainty about the value of X is reduced from $\log_2 6$ to $\log_2 3$ bit when knowing Y.

The joint information is related to the conditional information by

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{3}$$

Thus, $H(X,Y) \leq H(X) + H(Y)$, and $<$ if X and Y are not independent. In the example, we have $H(X,Y) = H(X)$, since the value of X determines that of Y.

The information gain about X from knowing Y is called the *mutual information* of X and Y,

$$MI(X:Y) = H(X) - H(X|Y). \tag{4}$$

In our example $MI(X:Y) = \log_2 6 - \log_2 3 = \log_2 2 = 1$ bit. From Y, we gain 1 bit of information about X.
The mutual information is symmetric,

$$MI(X:Y) = MI(Y:X). \tag{5}$$

The difference structure is perhaps the most important aspect. In many respects, (4) is more important and fundamental than (1), because we always have some prior knowledge, expressed here through Y, when we observe some X. Thus, the mutual information $MI(X:Y)$ tells us how much we can already infer about X when we know Y. By then observing X, we only gain the additional information $H(X|Y)$.

Summary:

$$H(X) = MI(X:Y) + H(X|Y) \tag{6}$$

$H(X)$ = how much you learn from observing X
$MI(X:Y)$ = how much you learn about X by observing Y
$H(X|Y)$ = how much you learn from observing X when you already know Y.

We can iterate the conditioning process with another random variable Z, to get the *conditional mutual information*

$$MI(X:Y|Z) = H(X|Z) - H(X|Y,Z). \tag{7}$$

$MI(X:Y|Z)$ quantifies how much additional mutual information between X and Y can be gained when we already know Z.
Careful: While always $H(X|Z) \leq H(X)$, we do *not necessarily* have $MI(X:Y|Z) \leq MI(X:Y)$.

Example: The XOR function (exclusive *or*):

x	y	z
0	0	0
1	0	1
0	1	1
1	1	0

where X, Y assume their two values independently with probability $1/2$ each. Thus, $MI(X:Y) = MI(X:Z) = MI(Y:Z) = 0$, but $MI(X:Y|Z) = MI(X:Z|Y) = MI(Y:Z|X) = 1$, because knowing the values of two of the variables determines that of the third.

2.3 Maximum entropy

E.Jaynes' maximum-entropy principle [Jaynes 2003]: Take the least informative estimate possible on the given information, that is, don't put any information into your model that is not based on the observed data. Look for p with maximal entropy $H(p)$ under the constraint that the expectation values of certain observables f_α be reproduced,

$$E_p f_\alpha = \sum_i f_\alpha^i p_i \text{ for } \alpha = 1, \ldots, A. \tag{8}$$

The solution is an exponential distribution

$$p_j = \frac{1}{Z} \exp(\sum_\alpha \lambda_\alpha f_\alpha^j) \quad \text{with } Z = \sum_i \exp(\sum_\alpha \lambda_\alpha f_\alpha^i). \tag{9}$$

In particular, when there are no observations,

$$p_j = \frac{1}{n} \text{ for } j = 1, \ldots, n. \tag{10}$$

2.4 Kullback-Leibler divergence

A reference for the information geometric concepts that will be introduced and used here and in the sequel is [Ay et al. 2017]. The *Kullback-Leibler divergence*

(*KL-divergence* for short) or *relative entropy* for two probability distributions p, q

$$D(p\|q) = \begin{cases} \sum_i p_i \log_2 \frac{p_i}{q_i} & \text{if supp } p \subset \text{supp } q \\ \infty & \text{else} \end{cases} \tag{11}$$

is positive ($D(p\|q) > 0$ if $p \neq q$), but not symmetric, as in general, $D(p\|q) \neq D(q\|p)$.

Example: The mutual information is the KL-divergence between the joint distribution and the product of the marginals,

$$MI(X:Y) = D(p(x,y)||p(x)p(y)). \tag{12}$$

Among all distributions $p(x,y)$ with the same marginals $p(x) = \sum_y p(x,y), p(y) = \sum_x p(x,y)$, the product distribution $p(x)p(y)$ has the largest entropy. This is, of course, a special case of Jaynes' principle. That is, when we only know the marginals, Jaynes' principle would suggest to take the product distribution as our estimate.

Example: The space of all probability distributions on two binary variables is a 3-dimensional simplex. It contains the 2-dimensional subfamily of product distributions. The extreme points of the simplex are the Dirac measures $\delta^{(x,y)}, x, y = 0, 1$. Maximization of the distance from the family of product distributions leads to distributions with support cardinality two (perfect correlation or anticorrelation) [Ay 2002].

The formal way of expressing Jaynes' principle is to project a given distribution onto the product family $\mathcal{E}$ to maximize entropy while preserving the marginals, with π denoting that projection,

$$\begin{aligned} D(p \,\|\, \mathcal{E}) \quad &:= \quad \inf_{q \in \mathcal{E}} D(p \,\|\, q) \;=\; D(p \,\|\, \pi(p)) \\ &= \quad H_{\pi(p)}(X,Y) - H_p(X,Y). \end{aligned} \tag{13}$$

3 Complexity

In this section, we want to introduce and discuss complexity concepts. But what is *complexity*? Some possible answers (see [Ay et al. 2011, Ay et al. 2017] for a systematic discussion): **Complexity** is

1. the minimal effort or the minimal resources needed to describe or generate an object. Examples of such complexity concepts include algorithmic complexity (Kolmogorov [Kolmogorov 1965], Chaitin [Chaitin 1966], Solomonoff [Solomonoff 1964]); computational complexities; entropy (Shannon [Shannon 1948]), or entropy rate (Kolmogorov [Kolmogorov 1965], Sinai [Sinai 1959]).

2. the minimal effort or the minimal resources needed to describe or generate the *regularities* or the *structure* of an object. Examples of such complexity concepts include Kolmogorov minimal sufficient statistics and

related notions, stochastic complexity (Rissanen [Rissanen 1989]), effective complexity (Gell-Mann and Lloyd [Gell-Mann and Lloyd 1996]), excess entropy ([Shaw 1984], also known as effective measure complexity [Grassberger 1986]), forecasting complexity ([Zambella and Grassberger 1988]), also introduced as statistical complexity by Crutchfield, Young, Shalizi [Crutchfield and Young 1989, Shalizi and Crutchfield 2001]).

3. the extent to which an object, as a whole, is more than the sum of its parts (Aristotle [Aristoteles Metaphysik]), that is, the extent to which the whole cannot be understood by the analysis of the parts of the system in isolation, but only by also considering their interactions.

In order to systematically explore these aspects, we start with the most basic concept, that of algorithmic complexity [Kolmogorov 1965, Solomonoff 1964, Chaitin 1966] (see [Li and Vitányi 1997] for a systematic exposition). This concept expresses 1) in its purest form.

3.1 Algorithmic complexity

The *algorithmic complexity* of an object, such as a number or a piece of text, is the length of the shortest computer program that generates or produces the object as output.[1] Typically, one cannot compute this complexity, but only provide an upper bound by producing a computer program, but does not know whether this is the shortest possible one.
From a conceptual perspective, the basic premise is that irregular or random structures have the highest algorithmic complexity, because they do not admit a short description. In other words, we want to characterize the complexity of a structure by the difficulty of its description. That is, we ask the question: How much can the description of a structure be simplified by utilizing regularities?

- Very simple structures need not be simplified any further.
- Random structures cannot be simplified.
- *Computational complexity* (see for instance the expositions in [Papadimitriou 2004] and [Moore and Mertens 2011]): Running time of shortest computer program that can generate the structure: A simple structure is produced quickly, whereas for a random one, everything has to be explicit in the program, and so, it does not need to run for a long time either.
- Random structures are not of interest for themselves, but only as members of an ensemble; it therefore suffices to describe the latter (Gell-Mann and Lloyd [Gell-Mann and Lloyd 1996]).

[1] to make the complexity of different objects comparable, one needs to agree on a predetermined programming language; usually, one assumes some universal Turing machine, and changing that Turing machine will introduce an additive constant in the upper bounds

3.2 External and internal complexity

The question that arises from the above concept of algorithmic complexity is how to compute it, that is, how to find the shortest description of a given structure. Quite apart from the fact that this depends on the choice of the device we use to evaluate it (in theory: some universal Turing machine, and the choice of that Turing machine then introduces an additive constant), in practice, we have only bounded means to represent a structure. Thus: What do we want to know? We want to

1. know a rich and complex structure,
2. but represent it most efficiently.

More formally, we want to

1. maximize *external complexity*,
2. but minimize *internal complexity*.

This perspective was introduced in [Jost 2004]. For an application in pattern classification, see for instance [Avdiyenko et al. 2015].

3.3 Optimization principles

Organisms live in and interact with a complex environment, see for instance [von Uexküll 2014] (for a measure theoretical approach, see [Ay and Loehr 2015]), and need to maintain their own autopoiesis [Maturana and Varela 1979]. A modern society consists of several complex subsystems that follow their own rules, but need to interact with each other [Luhmann 1984, Luhmann 1997]. With the concept of Shannon information, we can formulate some abstract principles that either maximize or minimize some kind of complexity (we follow [Jost 2016a] here). The basic versions, however, lead to trivial results, as we shall now see.

1. Gain as much information as possible: Look at random patterns
2. Avoid surprises: Look at blank screen
3. Try to predict future sensory inputs as accurately as possible on the basis of the current ones (and perhaps try to bring yourself into a state where this is possible [Friston 2010])
4. Try to manipulate the environment such that the results of own actions are as accurately predictable as possible [Klyubin et al. 2005].
5. Maximize

$$\begin{array}{rcl} & MI(S_{t+1}:E_t) & -MI(S_{t+1}:E_t|S_t) \\ = & H(S_{t+1}) - H(S_{t+1}|E_t) & -H(S_{t+1}|S_t) + H(S_{t+1}|E_t,S_t) \end{array} \tag{14}$$

to establish the strongest possible correlation between the current state E_t of the environment and future sensory data S_{t+1}, but such that this correlation can already be predicted from the current input S_t [Bertschinger et al. 2008]

To proceed further, let us discuss some questions.

1. Q: Why should a system model an external probability distribution?
   A: To make predictions on the basis of regularities
2. Q: How can this be achieved in an environment that is vastly more complex than the system itself?
   A: Detect regularities
3. Q: How to detect regularities?
   A: Because of 2), the system is forced to compress.

These answers have some consequences in various fields:

- Psychology: Use heuristics [Simon 1955, Simon 1956, Gigerenzer and Todd 1999]
- Cognition: External vs. internal complexity [Jost 2004]
- Statistics: Avoid overfitting
- Statistical learning theory: Start with models with few parameters and gradually increase as you learn (Vapnik-Chervonenkis) [Vapnik 1995, Vapnik 1998]

3.4 Correlations in time series

We can also use the information theoretical notions to evaluate the complexity of a time series in terms of the correlations that it exhibits. A time series $X_t, t \in \mathbb{N}$ could possibly have

- No regularities: $H(X_t|X_{t-1}) = H(X_t)$
- the Markov property: $H(X_t|X_{t-1}, X_{t-2}, \dots) = H(X_t|X_{t-1})$, or
- Long term correlations, as in texts, genetic sequences, ...

To evaluate this, we quantify how much new information is gained when one already knows n consecutive symbols and then sees the $(n+1)$st. (Grassberger [Grassberger 1986])
For which n is this largest? When n is small, one perhaps cannot predict much, and if n is large, one may be able to guess the rest anyway.
The larger this n, the more complex the sequence.
For genetic sequences, $n \sim 14$ [1], for amino acid sequences (proteins) $n \sim 5$.
In literature analysis, such a principle can be used to evaluate the complexity of language [Efer et al. 2015].

A more sophisticated concept is the genon concept of molecular biology [Scherrer and Jost 2007a, Scherrer and Jost 2007b, Jost and Scherrer 2014].

3.5 Complementarity

Instead of trying to predict the environment, one can also let the environment do the computation itself (see [Jost 2016a]).
If you want to catch a ball, you do not use Newtonian mechanics to compute the trajectory, but simply run so that the ball appears under a constant angle. The environment computes the trajectory, and you only need to sample. This outsourcing of computation represents one mechanism for the compression mentioned in Section 3.3.

More generally, embodied cognition has emerged as a new paradigm in robotics [Pfeifer and Bongard 2007].

3.6 Hierarchical models and complexity measures

In this section, we follow [Ay et al. 2011] and [Ay et al. 2017]. Returning to Jaynes' approach, we could maximize entropy while preserving marginals among subsets of variables. For instance, for a distribution on 3 variables, we could prescribe all single and pairwise marginals.

Assume that we have a state set V that consists of the possible values of N variables. We then consider the hierarchy

$$\mathfrak{S}_1 \subseteq \mathfrak{S}_2 \subseteq \ldots \subseteq \mathfrak{S}_{N-1} \subseteq \mathfrak{S}_N := 2^V, \tag{15}$$

where $\mathfrak{S}_k$ is the family of subsets of V with $\leq k$ elements, from wich we get the set of probability distributions $\mathcal{E}_{\mathfrak{S}_k}$ with dependencies of order $\leq k$. For instance, $\mathcal{E}_{\mathfrak{S}_1}$ is the family of distributions that are simply the products of their marginals. In particular, for a probability distribution in this family, there are no correlations between the probabilities of two or more of the variables. In $\mathcal{E}_{\mathfrak{S}_2}$, we then allow for pairwise correlations, but no triple or higher order ones. We point out that one can also consider other families of subsets of V and the corresponding probability distributions. For instance, when V is the ordered set of integers $\{1, \ldots, N\}$, one could consider the family of those subsets that consist of uninterrupted strings of length $\leq k$. This will be our choice when we discuss the excess entropy below.
We let $\pi_{\mathfrak{S}_k}$ be the projection on $\mathcal{E}_{\mathfrak{S}_k}$, $p^{(k)} := \pi_{\mathfrak{S}_k}(p)$. For instance, $p^{(1)}$ is the product distribution with the same marginals as p.
We have the important Pythagorean relation

$$D(p^{(l)} \,\|\, p^{(m)}) \;=\; \sum_{k=m}^{l-1} D(p^{(k+1)} \,\|\, p^{(k)}), \tag{16}$$

for $l, m = 1, \ldots, N-1$, $m < l$. In particular,

$$D(p \,\|\, p^{(1)}) \;=\; \sum_{k=1}^{N-1} D(p^{(k+1)} \,\|\, p^{(k)}). \tag{17}$$

If we take configurations with dependencies of order $\leq k$, we get the **Complexity measure** [Ay et al. 2011] with weight vector $\alpha = (\alpha_1, \ldots, \alpha_{N-1}) \in \mathbb{R}^{N-1}$

$$C_\alpha(p) := \sum_{k=1}^{N-1} \alpha_k \, D(p \,\|\, p^{(k)}) \tag{18}$$

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

with $\beta_k := \sum_{l=1}^{k} \alpha_l$.
$p^{(k)}$ is the distribution of highest entropy among all those with the same correlations of order $\leq k$ as p.
Thus, we consider a weighted sum of the higher order correlation structure.
Examples:

- Tononi-Sporns-Edelman complexity [Tononi et al. 1999]: $\alpha_k = \frac{k}{N}$ addresses the issue of the interplay between differentiation and integration in complex systems (for an analysis of system differentiation from an information theoretical perspective, see also [Jost et al. 2007])
- Stationary stochastic process X_n: Conditional entropy

  $$h_p(X_n) := H_p(X_n \,|\, X_1, \ldots, X_{n-1}).$$

  Entropy rate or *Kolmogorov–Sinai entropy* [Kolmogorov 1965, Sinai 1959]

  $$h_p(X) := \lim_{n\to\infty} h_p(X_n) = \lim_{n\to\infty} \frac{1}{n} H_p(X_1, \ldots, X_n), \tag{20}$$

  Excess entropy (Grassberger [Grassberger 1986])

  $$E_p(X) := \lim_{n\to\infty} \sum_{k=1}^{n} (h_p(X_k) - h_p(X))$$

  $$= \lim_{n\to\infty} (H_p(X_1, \ldots, X_n) - n h_p(X)) \tag{21}$$

  $$= \lim_{n\to\infty} \underbrace{\sum_{k=1}^{n-1} \frac{k}{n-k} \, D(p_n{}^{(k+1)} \,\|\, p_n{}^{(k)})}_{=: E_p(X_n)}, \tag{22}$$

  where we choose $\mathfrak{S}_k$ as the sequences of integers $j+1, j+2, \ldots, j+\ell$ with $\ell \leq k$. The excess entropy measures the non-extensive part of the entropy, i.e. the amount of entropy of each element that *exceeds* the entropy rate.

3.7 Interactions between levels

The question of emergence, that is, how a higher level that is (at least partially) autonomous from lower levels, arises in many disciplines. For example, classical

mechanics arises from an underlying quantum structure, but the laws of classical mechanics are causally closed, in the sense that for computing trajectories of Newtonian particles, we do not need information from the quantum level. Likewise, human genetics rests on the laws of Mendel and does not need to consider an underlying biochemical level. In other fields it is often not so clear, however, to what extent laws operate autonomously at a certain level without needing permanent or at least regular access to some lower level. For instance, does it suffice for understanding macroeconomic processes to consider relations between macroeconomic variables, or is an input from the microeconomic level essentially needed? Or can one understand social dynamics without access to the psychic and mental states of the participating individuals? For a general discussion of the issue of emergence from the perspective developed in the present contribution, see for instance [Jost et al. 2010].

Here, we describe the approach of [Pfante et al. 2014a, Pfante et al. 2016] (and refer to [Pfante et al. 2014a] for references to earlier work). We consider a structure

$$\begin{array}{ccc} \widehat{X} & \xrightarrow{\psi} & \widehat{X}' \\ {\scriptstyle\pi}\uparrow & & \uparrow{\scriptstyle\pi} \\ X & \xrightarrow{\phi} & X' \end{array}$$

with basic level X, X' and higher level $\widehat{X}, \widehat{X}'$; an arrow $Y \to Y'$ represents a discrete time step where X, X' form a Markov process, with transition kernel ϕ, which can be observed at the higher level $\widehat{X}, \widehat{X}'$ in a lossy fashion.

The higher level could result from averaging or aggregating the lower level. Think of $\widehat{X}$ as a coarse-graining of X given by an observation map π.

We can propose several criteria for the upper process being closed in the sense that it depends on the lower process only through some initialization.

I **Informational closure**: The higher process is informationally closed, i.e. there is no information flow from the lower to the higher level. Knowledge of the microstate will not improve predictions of the macrostate.

$$MI(\widehat{X}' : X|\widehat{X}) = 0 \tag{23}$$

where the conditional mutual information

$$MI(\widehat{X}' : X|\widehat{X}) = H(\widehat{X}'|\widehat{X}) - H(\widehat{X}'|X) \tag{24}$$

measures the reduction in uncertainty about $\widehat{X}'$ when knowing X instead of only $\widehat{X}$.

II **Observational commutativity**: It makes no difference whether we perform the aggregation first and then observe the upper process, or we observe the process on the microstate level, and then lump together the

states.
Kullback-Leibler divergence between the lower and the upper transition kernel from X to $\widehat{X}'$ is 0, for some initial distribution on X.

$$I \Rightarrow II, \text{ and in deterministic case also } II \Rightarrow I. \tag{25}$$

(In I, probabilities at $\widehat{X}$, in II at X)

III **Commutativity**: There exists a transition kernel ψ such that the diagram commutes (Görnerup-Jacobi, 2010)

$$II \Rightarrow III, \text{ and in deterministic case also } III \Rightarrow II. \tag{26}$$

II: Transition kernels satisfy $\Psi = \Pi\Phi\Pi^T$
III: Transition kernels satisfy $\Psi\Pi = \Pi\Phi$

IV **Markovianity**: $\widehat{X}, \widehat{X}'$ forms again a Markov process (Shalizi-Moore, 2003).

$$I \Rightarrow IV, \text{ but } IV \nRightarrow III. \tag{27}$$

V **Predictive efficiency**: A more abstract formulation is that an emergent level corresponds to an efficiently predictable process, that is, one that can be predicted in its own terms, without permanent recourse to a lower level.

3.7.1 A test case: The tent map

We now evaluate the preceding concepts at the example of the tent map, following [Pfante et al. 2014b] (see also [Atay et al. 2009] for background).

$$T(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1/2 \\ 2 - 2x & \text{else} \end{cases}$$

The tent map is a basic example of a chaotic dynamical iteration, because at every step differences between values can get doubled, and therefore, after several steps, even very tiny differences between initial values can become macroscopically large. The folding at $x = 1/2$ ensures that nevertheless the unit interval is mapped to itself. Thus, some differences also get reduced. Understanding this interplay between amplification and reduction of differences is surprisingly subtle, as one may also see in the following.

For a threshold value $\alpha \in [0, 1]$ we define the symbolic dynamics

$$\begin{aligned} \phi_\alpha : \quad & X \to \hat{X} = \{0, 1\} \\ \phi_\alpha(x) \quad := & \begin{cases} 0 & \text{if } 0 \leq x < \alpha \\ 1 & \text{else} \end{cases} \end{aligned}$$

The sequence $x_n = T^n(x)$, for an initial value $x \in X$, yields the derived symbol dynamics $s_n = \phi_\alpha(x_n) \in \{0, 1\}$.

The probability of finding s_n in the state 0 is the probability that x_n lies in the interval $[0, \alpha]$ (which is α for the tent map).

We consider the symbolic dynamics derived from consecutive time steps

$$(s_{n+m}, s_{n+m-1}, \ldots, s_n)\,,$$

with $k \in \mathbb{N}$

$$s_k(x) = \begin{cases} 0 & \text{if } T^k(x) < \alpha \\ 1 & \text{if } T^k(x) \geq \alpha \end{cases} .$$

For comparison, we take a random sequence $\xi_n \in [0, 1]$ (uniformly, i.i.d.), and consider the corresponding symbolic dynamics

$$\sigma_n := \begin{cases} 0 & \text{if } \xi_n < \alpha \\ 1 & \text{if } \xi_n \geq \alpha. \end{cases}$$

The question now is: Are there systematic differences between the symbolic sequence s_n derived from iterations of the tent map and σ_n?

For $\alpha = 1/2$, they look the same (in fact, we simply have a Bernoulli sequence: the values 0 and 1 occur with equal probability 1/2; $p(0) = p(1) = 1/2$). If we don't know x, s_n looks as random as σ_n. The transition probabilities are

$$p(0|0) = p(1|0) = p(0|1) = p(1|1) = 1/2.$$

We next consider $\alpha = 2/3$. Put $x_n := T^n(x)$.
$\sigma_n = 0$ and $\sigma_n = 1$ occur independently with probabilities 2/3 and 1/3.

When $s_n = 1$, that is, $2/3 < x_n \leq 1$, then $0 \leq x_{n+1} < 2/3$, that is $s_{n+1} = 0$. Thus, there is no transition from 1 to 1. For the state $s_n = 0$, both transitions are equally likely: when $0 \leq x_n \leq 1/3$, we have $0 \leq x_{n+1} \leq 2/3$, that is, $s_{n+1} = 0$, while for $1/3 < x \leq 2/3$, we get $s_{n+1} = 1$. Thus, for s_n,

$$p(0|0) = p(1|0) = 1/2,\ p(0|1) = 1,\ p(1|1) = 0$$

while for σ_n

$$p(0|0) = p(0|1) = 2/3,\ p(1|0) = p(1|1) = 1/3.$$

This leads us to the concept of *forbidden sequences*. While for the threshold $\alpha = 1/2$, the symbolic dynamics of the tent map cannot be distinguished from that of a random sequence, and is Markovian, in contrast, for the threshold $\alpha = 2/3$, the sequence 11 does not occur, and the symbolic dynamics is different from a random one, but still Markovian.
For other thresholds, we can also get longer forbidden sequences and non-Markovian symbolics.

Even from a random sequence ξ_n, we can derive non-Markovian symbolic dynamics.
Let $x^1, x^2 \in [0,1]$; we consider the symbolic rule

$$s(x^1, x^2) = \begin{cases} 0 & \text{if } x^1 \leq x^2 \\ 1 & \text{if } x^2 < x^1. \end{cases}$$

For our random sequence, take $x^1 = \xi_n, x^2 = \xi_{n+1}$. Thus, we draw the points x^1, x^2 randomly and independently.
The state probabilities are again $p(0) = p(1) = 1/2$, but the transition probabilities now depend on the history. The more 1s we have seen, the less likely it is to see another 1, because then ξ_n is expected to be very small, hence most likely, $\xi_{n+1} > \xi_n$.

We now analyze the *information flow* of this example. The information flow between the micro-level corresponding to state x_n and the coarse-grained level s_n is the conditional mutual information

$$MI(s_{n+1} : x_n | s_n) = H(s_{n+1}|s_n) - H(s_{n+1}|s_n, x_n)\,.$$

Since s_{n+1} is fully determined by x_n, the second term vanishes,

$$MI(s_{n+1} : x_n | s_n) = H(s_{n+1}|s_n)\,,$$

i.e., the information flow = conditional entropy on the coarse grained level, which has a local minimum at $\alpha = 2/3$.

Instead of drawing information from below, the upper level system relies on its memory.

4 Information decomposition

We finally turn to the concept of information decomposition. To motivate it, we start with the *transfer entropy* [Schreiber 2000][2]

$$TE(Z \to X) := MI(X_+ : Z_- | X_-) \tag{28}$$

where the subscript $-$ refers to the past and $+$ to the future. $TE(Z \to X)$ quantifies the amount of information contained in Z about the future of X that cannot be obtained from its own past.
Problem: $X_+ = \mathrm{XOR}(X_-, Z_-)$:
Here, the information in Z_- is only useful together with that of X_-. The transfer entropy cannot distinguish this situation from one where X_- does not

[2]Such a principle had already been introduced by the econometrician Granger [Granger 1969] who wrote "We say that Y_t is causing X_t if we are better able to predict X_t using all available information than if the information apart from Y_t had been used." In the econometric literature, this principle was applied only in linear settings. As [Barnett et al. 2009], explained, the transfer entropy can be seen as an operationalization of this principle in a general context.

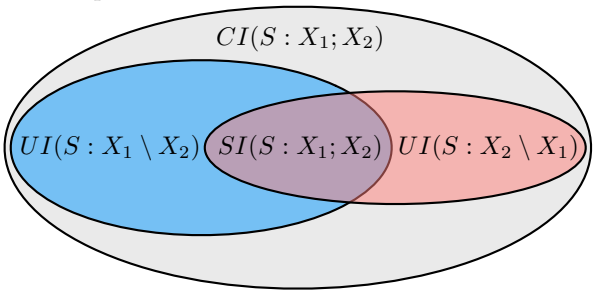
contribute and Z_- determines X_+ by itself.

This problem is addressed by information decomposition. It was started by Williams and Beer [Williams and Beer 2010] (but their measure $I_{\min}$ of shared information does not distinguish whether different random variables carry the same information or just the same amount of information), and continued by Harder, Salge, Polani [Harder et al. 2013], Griffith and Koch [Griffith and Koch 2014], Bertschinger, Rauh, Olbrich, Ay, Banerjee, Jost [Bertschinger et al. 2012, Bertschinger et al. 2014, Rauh et al. 2014, Rauh et al. 2017], and taken up by many other people (see for instance the references in [Lizier et al. 2018]), with applications in different fields, like neuroscience [Wibral et al. 2017]. There is no optimal solution, but that of Bertschinger, Rauh, Olbrich, Jost, Ay [Bertschinger et al. 2014] (called the *BROJA* decomposition in the community) is currently the most widely accepted.

To describe our approach, we consider three random variables X_1, X_2 and S. The (total) mutual information $MI(S : X_1, X_2)$ quantifies the total information that is gained about S if the outcomes of X_1 and X_2 are known. How do X_1 and X_2 contribute to this information? For two explanatory variables, we expect four contributions to $MI(S : X_1, X_2)$:

$$\begin{aligned} MI(S : X_1, X_2) \;=\; & SI(S : X_1; X_2) && \text{shared information} \\ & +UI(S : X_1 \backslash X_2) && \text{unique information of 1} \\ & +UI(S : X_2 \backslash X_1) && \text{unique information of 2} \\ & +CI(S : X_1; X_2) && \text{complementary or synergistic information.} \end{aligned}$$

Here, $UI(S : X_1 \backslash X_2)$ is the information that X_1 has, but X_2 does not have, $SI(S : X_1; X_2)$ is the information that both of them have individually. Perhaps the most interesting term is the last, $CI(S : X_1; X_2)$, the information that only emerges if X_1 and X_2 pool their knowledge. This term is best illustrated in the **XOR** example discussed below.



Gray: $MI(S : X_1, X_2)$
Blue: $MI(S : X_1)$
Orange: $MI(S : X_2)$

We consider some examples. **AND**

x_1	x_2	s	$p(x_1, x_2, s)$
0	0	0	1/4
1	0	0	1/4
0	1	0	1/4
1	1	1	1/4

.

Here, x_1 and x_2 jointly determine s, but cannot be fully recovered from s.

When 1 has the value $x_1 = 0$, she can exclude $s = 1$, and analogously for 2. Thus, when they both see 0, they share the information that $s = 0$.

The mechanism loses some information. When X_1, X_2 are i.i.d.,

$$H(X_1, X_2) = 2 \text{ bits},$$

but

$$H(S) = MI(S : X_1, X_2) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} \approx .811 \text{ bits}.$$

In general, we may have both correlations between the input variables and relations created by the mechanism that computes S.

We next recall **XOR** from Section 2.2:

x_1	x_2	s
0	0	0
1	0	1
0	1	1
1	1	0

.

Neither 1 nor 2 can determine the value of S by herself, but the value of the other is needed for that. This is a clear case of synergistic information only.

Our approach: Unique and shared information should only depend on the marginal distribution of the pairs (S, X_1) and (S, X_2). This idea can be explained from an operational interpretation of unique information: Namely, if X_1 has unique information about S (with respect to X_2), then there must be some way to exploit this information. More precisely, there must be a situation in which X_1 can use this information to perform better at predicting the outcome of S.

In this interpretation, 1 possesses unique information about S compared with 2, if there exists a reward function for which 1 can achieve a higher expected reward based on her value x_1 and her knowledge of the conditional distribution $p(s|x_1)$ than if she knew and utilized instead the conditional distribution of 2.

Thus, unique and shared information depend only on pairwise marginals. Only the synergistic information includes higher order dependencies. In that sense, synergy becomes a measure of higher order interactions, in the sense of information geometry.

From a conceptual perspective, and independently of the way the different terms in the decomposition are quantified, it is important to understand synergy, in order to clarify discussions that have become quite sterile, like the relative importance of genes and environment in biology. For a perspective in this direction, see [Jost 2020].

References

[Aristoteles Metaphysik] Aristoteles, Philosophische Schriften 5. Metaphysik. Nach der Übersetzung von Hermann Bonitz bearbeitet von Horst Seidl, Felix Meiner, Hamburg, 1995

[Atay et al. 2009] F. Atay, S. Jalan und J. Jost, Randomness, chaos, and structure, Complexity 15, 2009, 29–35

[Avdiyenko et al. 2015] Avdiyenko, L., N. Bertschinger und J. Jost. 2015. Adaptive Information-Theoretical Feature Selection for Pattern Classification. Computational Intelligence. Springer International Publishing, 279-294.

[Ay 2002] N. Ay, An Information-geometric approach to a theory of pragmatic structuring, Ann. of Prob., v. 30, N. 1 (2002) 416–436.

[Ay et al. 2017] N.Ay, J.Jost, H.V.Lê, L.Schwachhöfer, *Information geometry*, Springer, Ergebnisse der Mathematik, 2017

[Ay et al. 2011] N.Ay, E.Olbrich, N.Bertschinger und J.Jost, A geometric approach to complexity, Chaos 21, 037103 (2011); doi: 10.1063/1.3638446

[Ay and Loehr 2015] N.Ay und W.Loehr, The Umwelt of an Embodied Agent A Measure-Theoretic Definition, Theory Biosc. 2015

[Barnett et al. 2009] L. Barnett, A. Barnett, A.Seth, Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables, Phys. Rev. Lett. 103 (23), 38701, 2009

[Bertschinger et al. 2008] Bertschinger, N., E. Olbrich, N. Ay und J. Jost. 2008. Autonomy: An information theoretic perspective. *Biosystems* **91**: 331–345

[Bertschinger et al. 2012] Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. *Shared information – new insights and problems in decomposing Information in Complex Systems* , Proc. ECCS 2012, 251–269

[Bertschinger et al. 2014] Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.: Ay, N. *Quantifying Unique Information*, Entropy 2014, 16, 2161-2183; doi:10.3390/e16042161

[Bertschinger et al. 2012] Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared InformationNew Insights and Problems in Decomposing Information in Complex Systems. In Proceedings of the European Conference on Complex Systems 2012, Brussels, Belgium, 27 September 2012; pp. 251269.

[Bertschinger et al. 2014] Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. Entropy 2014, 16, 21612183.

[Chaitin 1966] G.Chaitin, On the lengths of programs for computing finite binary sequences, JACM 13, 547-569 (1966)

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

[Efer et al. 2015] T.Efer, G.Heyer und J.Jost, in: Shakespeare unter den Deutschen (C.Jansohn, Hrsg.)

[Friston 2010] Friston, K. 2010. The free-energy principle: a unified brain theory? *Nature Review Neurosc.***11**:127-138

[Gell-Mann and Lloyd 1996] M.Gell-Mann, S.Lloyd, Information measures, effective complexity, and total information, Complexity 2, 44-52 (1996)

[Gigerenzer and Todd 1999] Gigerenzer, G., und P.Todd. 1999. Simple heuristics that make us smart, Oxford University Press

[Granger 1969] C.Granger, Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica Vol. 37 (1969), 424–438

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

[Griffith and Koch 2014] Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In Guided Self-Organization: Inception; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germary, 2014; Volume 9, pp. 159190.

[Harder et al. 2013] Harder, M.; Salge, C.; Polani, D. A Bivariate measure of redundant information. Phys. Rev. E 2013, 87, 012130.

[Jaynes 2003] E.T.Jaynes, Probability theory: The logic of science, Cambridge Univ.Press, 2003

[Jost 2004] Jost, J. 2004. External and internal complexity of complex adaptive systems, *Theory Biosc.***123**: 69–88

[Jost 2016a] Jost, J. 2016. Sensorimotor contingencies and the dynamical creation of structural relations underlying percepts. In: Proc. Strüngmann Forum: Wheres the Action? The Pragmatic Turn in Cognitive Science, MIT Press, im Druck

[Jost 2020] Jost, J. 2020, Biological information, to appear in Theory in Biosciences

[Jost et al. 2007] J. Jost, N. Bertschinger, E. Olbrich, N. Ay und S. Frankel, An information theoretic approach to system differentiation on the basis of statistical dependencies between subsystems, Physica A 378, 2007, 1–10

[Jost et al. 2010] J. Jost, N. Bertschinger und E. Olbrich, Emergence. A dynamical systems approach, New Ideas Psych.28, 2010, 265–273

[Jost and Scherrer 2014] J.Jost und K.Scherrer, Information theory, gene expression, and combinatorial regulation - A quantitative analysis, Theory Biosc.133, 2014, 1–21

[Kolmogorov 1965] A.N.Kolmogorov, Three approaches to the quantitative definition of information, Problems of Information Transmission 1, 4-7 (1965)

[Klyubin et al. 2005] Klyubin, A., D. Polani und C. Nehaniv. 2005. Empowerment: A Universal Agent-Centric Measure of Control. In: Proc. IEEE CEC 2005: 128–135

[Li and Vitányi 1997] M.Li und P.Vitányi, An introduction to Kolmogorov complexity and its applications, Springer, ^{2}1997

[Lizier et al. 2018] Lizier, J., Bertschinger, N., Jost, J. and Wibral, M.. Information decomposition of target effects from multi-source interactions: perspectives on previous, current and future work. Entropy 2018, 20, 307; doi:10.3390/e20040307

[Luhmann 1984] N.Luhmann, Soziale Systeme, Suhrkamp, Frankfurt, 1984

[Luhmann 1997] N.Luhmann, Die Gesellschaft der Gesellschaft, 2 Bde., Suhrkamp, Frankfurt, 1997

[Maturana and Varela 1979] H.Maturana und. F.Varela, Autopoiesis and Cognition, Reidel, Boston, 1979

[Moore and Mertens 2011] C.Moore and S.Mertens, The nature of computation, Oxford Univ.Press, 2011

[Papadimitriou 2004] C. Papadimitriou, Computational Complexity, Addison Wesley, 1994.

[Pfante et al. 2014b] O. Pfante, E. Olbrich, N. Bertschinger, N. Ay, and J. Jost. Closure measures for coarse-graining of the tent map. Chaos, 24:013136, 2014.

[Pfante et al. 2014a] O. Pfante, N. Bertschinger, E. Olbrich, N. Ay, J. Jost, Comparison between different methods of level identification, Advances in Complex Systems 17, 2014, 1450007 (21 Seiten)

[Pfante et al. 2016] O. Pfante, N. Bertschinger, E. Olbrich, N. Ay, J. Jost, Wie findet man eine geeignete Beschreibungsebene für ein komplexes System?, in: Jahrbuch der Max-Planck-Gesellschaft, im Druck, 2016

[Pfeifer and Bongard 2007] R. Pfeifer und J. Bongard. How the body shapes the way we think. MIT Press, 2007

[Rauh et al. 2017] J.Rauh, P. Banerjee, E. Olbrich, J. Jost, and N. Bertschinger. On extractable shared information. En- tropy, 19(7):328, 2017.

[Rauh et al. 2014] Rauh, J.; Bertschinger, N.; Olbrich, E.; Jost, J. Reconsidering unique information: Towards a multivariate information decomposition. In Proceedings of 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June4 July 2014; pp. 22322236.

[Rissanen 1989] J. Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific 1989.

[1] M.Sadovsky, Genes, information and sense: complexity and knowledge retrieval, Theory in Biosciences 127(2), pp 69–78, 2008

[Scherrer and Jost 2007a] K.Scherrer und J.Jost, The gene and the genon concept: A functional and information-theoretic analysis. Mol Syst Biol 3:87, Epub 2007 Mar 13: EMBO and Nature Publishing Group.

[Scherrer and Jost 2007b] K.Scherrer und J.Jost, Gene and genon concept: Coding versus regulation, Theory Biosc. 126, 65-113, 2007

[Schreiber 2000] T.Schreiber, Measuring information transfer, PRL 85, 461–464, 2000

[Shalizi and Crutchfield 2001] C.R. Shalizi and J.P. Crutchfield, Computational mechanics: Pattern and prediction, structure and simplicity, Journal of Statistical Physics 104 (2001) 817–879.

[Shannon 1948] C.Shannon, The mathematical theory of communication, 1948, reprinted in: C.Shannon, W.Weaver, The mathematical theory of communication, ed. R. Blahut, B.Hajek, pp.29-125, Univ. Illinois Press, 1998

[Shaw 1984] R. Shaw, The dripping faucet as a model chaotic system, Aerial Press, Santa Cruz, 1984

[Simon 1955] Simon, H. A. 1955. A behavioral model of rational choice. Quart. J. Econom. 69(1) 99118.

[Simon 1956] Simon, H. A. 1956. Rational choice and the structure of environments. Psych. Rev. 63 129138

[Sinai 1959] Ja. Sinai, On the concept of entropy for a dynamical system, (Russian) Dokl. Akad. Nauk SSSR 124 (1959) 768–771.

[Solomonoff 1964] R.Solomomoff, A formal theory of inductive inference, Part I, Information and Control 7, 1-22, Part II, Information and Control 7, 224-254 (1964)

[Tononi et al. 1999] G. Tononi, O. Sporns, G. M. Edelman, Measures of degeneracy and redundancy in biological networks, PNAS 96 (1999) 3257–3267

[Vapnik 1995] Vapnik, V.N. 1995. The nature of statistical learning theory. Springer

[Vapnik 1998] V.N.Vapnik, Statistical learning theory, Wiley-Interscience, 1998

[von Uexküll 2014] J.J. von Uexküll, Umwelt und Innenwelt der Tiere, hrsg. v. F.Mildenberger, B.Herrmann, in: Klassische Texte der Wissenschaft, Springer Spektrum, 2014

[Wibral et al. 2017] Wibral, M.; Priesemann, V.; Kay, J.W.; Lizier, J.T.; Phillips, W.A. Partial information decomposition as a unified approach to the specification of neural goal functions. Brain Cogn. 2017, 112, 2538.

[Williams and Beer 2010] Williams, P.; Beer, R. *Nonnegative decomposition of multivariate information*, arXiv:1004.2515v1, 2010.

[Zambella and Grassberger 1988] D. Zambella and P. Grassberger, Complexity of Forecasting in a Class of Simple Models, Complex Systems 2, 269 (1988).