Equivalence of information production and generalised entropies in complex processes

Rudolf Hanel IDs, Stefan Thurner IDs*

1 Section for Science of Complex Systems, CeMDS, Medical University of Vienna, Vienna, Austria, 2 Complexity Science Hub Vienna, Vienna, Austria, 3 Santa Fe Institute, NM, Santa Fe, NM, United States of America

* These authors contributed equally to this work.

rudolf.hanel@meduniwien.ac.at

Abstract

Complex systems with strong correlations and fat-tailed distribution functions have been argued to be incompatible with the Boltzmann-Gibbs entropy framework and alternatives, so-called generalised entropies, were proposed and studied. Here we show, that this perceived incompatibility is actually a misconception. For a broad class of processes, Boltzmann entropy —the log multiplicity— remains the valid entropy concept. However, for non-i.i.d. processes, Boltzmann entropy is not of Shannon form,

\[-k \sum p_i \log p_i,
\]

but takes the shape of generalised entropies. We derive this result for all processes that can be asymptotically mapped to adjoint representations reversibly where processes are i.i.d. In these representations the information production is given by the Shannon entropy. Over the original sampling space this yields functionals identical to generalised entropies. The problem of constructing adequate context-sensitive entropy functionals therefore can be translated into the much simpler problem of finding adjoint representations. The method provides a comprehensive framework for a statistical physics of strongly correlated systems and complex processes.

Introduction

To know the information content of a process, a system, a source, a signal, or a sequence, one uses entropy to quantify it. If systems or processes are independent identically distributed (i.i.d.), ergodic and stationary in their probabilities, it is known what to do: one uses the expression [1],

\[ H(p) = - \sum_{i=1}^{W} p_i \log p_i , \tag{1} \]

where \( i = 1, 2, \ldots \), \( W \) are the states the system can take and \( p_i \) is the probability to observe them. So-called Shannon entropy is given by \( S(p) = k H(p) \), where \( k > 0 \) is a constant that specifies the units of entropy, in statistical physics \( k = k_B \) is the Boltzmann constant; if we measure information in bits per symbol, \( k = 1/\log 2 \). If the system or process of interest is not i.i.d., ergodic, or in stationary equilibrium, then it becomes less clear what to do in order to obtain its correct information content. In principle there are two conceptually very different paths to solve the problem.
The first way (which we present here in this paper) is to look at the information production of the process or system. Consider a process, $X$, that emits signals $x(T)$ of length $T$ with $x(T) = x_T x_{T-1} \ldots x_1$. Such an ordered list/sequence of elements $x(T)$ is often referred to as a $T$-tuple (or $T$-gram). Every symbol, $x_i$, in the signal is an element of a fixed sample space or an alphabet, $\mathcal{A}$, that contains all possible states $x_i$ can take. For example, if you think of $X$ as a text-producing author, states are the letters from the English alphabet, $\mathcal{A}^\text{letters}$, or in a “binary alphabet”, $\mathcal{A}^{0,1}$ of zeros and ones, once the text is stored on a computer. In non-i.i.d. systems, symbols within sequences will in general be correlated in one way or another. Those correlations—that may extend over many different scales in the systems—carry information about the system. Using the marginal distribution functions of the occurrences of states (letters) in Eq (1) will then certainly not provide the correct information content of the process. However, if we know the probability distribution, $p_T$, to observe entire sequences, $x(T)$, we compute the information production of process, $X$, as

$$I(X) = \lim_{T \to \infty} \frac{1}{T} kH(p_T), \tag{2}$$

see also Text 2 in S1 File for details. This changes the perspective from individual symbols, events, or states to entire sequences, or paths. From an information theoretic point of view, $I(X)$ measures the average number of bits required to reversibly encode samples of $X$ into bit-streams that can be sent through information channels and measure information production in bits per emitted symbol, i.e., $k = 1/\log(2)$. Note that $kH(p_T)$ measures bits per $T$-tuple, i.e. per path-segment-of-length-$T$. Therefore, by using $k_T = k/T$, one measures again bits per emitted symbol of the original alphabet. However, the number of all possible $T$-tuples—the size of the new “alphabet” is enormous if $T$ is large. If the sample space contains $A = |\mathcal{A}|$ symbols, the size of the alphabet for all paths is $A^T$. The sum over all states and knowing their probabilities, $p_T$, will in general be impossible.

The reason why $I(X)$ is the true information production rate of process $X$, is because there exist representations of $X$ in terms of other processes, $Y$, such that sequences, $x$, presented in the symbols from the initial alphabet, $\mathcal{A}_x$, can be rewritten into sequences, $y$, using other symbols from a much larger alphabet. The definition of information production, $I(X)$, asymptotically uses the largest alphabet containing each possible sequence as a unique symbol, which—as a consequence— are statistically independent. In other words, all structure (correlations) gets absorbed into new symbols belonging to an extended alphabet. Only, the definition of information production uses the largest possible alphabet; the alphabet of $T$-tuples over the original alphabet and $T \to \infty$.

However, typically, there exist much smaller alphabets that can capture all structures of a process or system. We say $Y$ is adjoined to $X$ (see subsection Constructing adjoint process spaces for details) if $Y$s symbols are uncorrelated and, thus, $H$ measures the correct information content. In general, the symbols, $z$, in the extended alphabet, encode for a number, $\ell$, symbols in the original alphabet. For instance, if $z$ is a symbol in the alphabet of $T$-tuples, then $T = z(\ell)$, $z$ consisting exactly of $T$ letters of the original alphabet. The average length, $\bar{\ell} = \langle \ell \rangle = \sum_z p_z(\ell(z))$, of original symbols emitted per symbol in the extended alphabet increases with the size of the extended alphabet, $f_z = p_z(Y)$ is the distribution function of letters $z$. Consequently, the unit of information, $k$, adapts to the “complexity” $\ell(z)$ and $k \rightarrow k/\bar{\ell}$.

For an example of how one can encode information about correlations of a process $X$ on all relevant scales imagine an initial alphabet of Latin letters and extend it to a series of extended alphabets: one that contains syllables in addition to letters, one that adds word-fragments, one that includes words, one with frequent word combinations, one with phrases, and so on. This
sequence of alphabets is nested in the sense that they contain each other, \( \mathcal{A}_0 \subset \mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots \subset \mathcal{A}_n \cdots \). With any one of these alphabets, say \( \mathcal{A}_n \), one can sample “text” from it by using the associated marginal distributions, \( p_\lambda(n) \), of its elements, \( \lambda \). With increasing \( n \), the resulting artificial text samples will more and more resemble the English text body from which the marginal distributions \( p_\lambda(n) \) were derived; see Text 1 in S1 File for the examples given by C. Shannon. We can find particular sequences of alphabets such that each alphabet, \( \mathcal{A}_{n+1} \), contains exactly one more symbol than \( \mathcal{A}_n \), by applying reversible substitutions of symbols, which we refer to as “parsing rules”. For details, see Text 3 in S1 File.

The second way is to directly capture the correlations and structures in non-i.i.d. systems / processes in a generalised functional form of the entropy, which –typically– looks more complicated than Eq (1), an approach that so far has been successful for a handful of processes. Generalised entropy functionals are usually expressed in terms of marginal distributions in the “original alphabet”. These generalised entropies have been extensively studied for several decades [2–6] from different angles, generally for systems with strong or long-range correlations [7–9], that are non-ergodic, internally constrained [10–12], or for systems out of equilibrium [13–15].

For non-i.i.d. systems or processes it is essential to specify the context in which the term entropy is used, whether one talks about information theory, thermodynamics, or the maximum entropy principle (MEP) [16]. While thermodynamic aspects of such systems, especially the existence of well defined thermodynamic potentials (and as a consequence, temperature) is heavily debated, there is wide consensus that entropy production (the physical analogue of information production) remains a valid concept, also for these systems. For thermodynamic considerations away from equilibrium and i.i.d. and applications e.g. to neural correlations see for instance [17–19].

Here we will not focus on thermodynamic aspects of entropy, but on the original context envisioned by Boltzmann: its power to predict a particular macro state from knowing the number of micro states (multiplicity) corresponding to it. This allows one to predict typical distribution functions and derive functional relations between macro state variables (expectation values of respective observables). This view that is tightly related to the MEP is not restricted to physics and is not limited to i.i.d. processes. For specific cases the respective MEP functionals, the generalised entropy and cross entropy, have been explicitly constructed [5, 6]. In particular, if multiplicity and probabilities are multiplicatively separable in the asymptotic limit, [6], a clear definition of cross entropy is possible also for non-i.i.d. systems.

While the Boltzmann entropy concept remains untouched (log multiplicity) its functional form, i.e., the generalised entropy functional, depends on the context of the process, \( X \), and the process class, \( \Phi \), to which it belongs to. Some examples for different process classes include i.i.d. processes, exchangeable and polynomial mixture processes [20], Polya processes [21], sample space reducing processes [6], and processes describing structure forming systems [22].

The idea behind generalised entropies is to quantify entropy as the logarithm of the number of micro states (multiplicity) of process, \( X \). It is based on the marginal distribution function, \( g_i = p_i(X) \), of symbols \( i \) from a sample space (here the original alphabet is \( \mathcal{A}_0 \)) that compose a functional, \( \tilde{S}(g) \), such that it captures the structural information in \( X \). Similarly, one obtains generalised expressions for cross entropy and information divergence. Those functionals effectively capture arbitrarily complicated relations between symbols of the sample space (original alphabet) in terms of marginal symbol frequencies, \( g_i \), of the process \( X \). Generalised entropies (that fulfil the first Shannon-Khinchin axiom) do not explicitly depend on system parameters that identify a process within a process class or other details. It is obvious that in general, constructing such a functional may be complicated and has not been achieved convincingly,
except for a few exceptions, e.g. [6, 21]. As we will see, one can reconstruct (or at least approximate) such functionals from data—at least in principle, since there exist fundamental limits to reconstructing generative grammars from data on the basis of statistical inference alone, [23, 24], a fact also captured in Chaitin’s incompleteness theorem, [25]. In other words, the question of whether some data, a particular sequence, \( x \), contains regular structures that can be used to compress it, may become undecidable.

Here we show that the two approaches, information production and the generalised entropy functionals can be mapped to one another, meaning that they are the same. The diagram in Fig 1 schematically shows the basic idea: We first use the method of parsing rules to construct an adjoint representation, \( Y \), of a given process \( X \), and write \( Y = \pi X \). Here \( \pi \) is a map that reversibly encodes all structures in \( X \), such that process \( Y \) in its new extended alphabet is i.i.d. and is therefore fully described by their marginal distribution function, \( f_z = p_z(Y) \), where \( z \) is again a letter from the extended alphabet. Consequently, the Shannon information measure with the appropriate unit of information is adequate (first way). Next, we project the marginal distributions from the adjoint representation, \( f_z \), to the original alphabet and, in a last step, we identify the process specific “pull-back” information measures, \( S_X, S^\text{cross}_X \), and \( D_X \), which get precisely defined in Eq (14) below, that takes distribution functions over the original alphabet, with the corresponding generalised information measure (second way) by adequately lifting distribution functions over the original alphabet to distributions over the adjoint alphabet.

![Diagram](https://doi.org/10.1371/journal.pone.0290695.g001)

**Fig 1.** Diagram of the relations between of distribution functions and entropies over the sample space and adjoint samples space. We consider a process, \( X \) over alphabet \( A \) with adjoint process \( Y = \pi X \) over the adjoint alphabet \( A^* \). \( Y \) is i.i.d. and therefore fully characterized by the marginal distribution of letters it samples from, i.e. asymptotically data \( y = \pi x \) is fully characterized by the relative frequency distribution function \( f = p(y) \). \( \Phi^* \) is the set of all i.i.d. processes over \( A^* \). Therefore \( \Phi^* = \pi^{-1} \Phi \) is the class of processes that naturally generalize \( X \in \Phi_a. \) can be projected to the marginal distribution function \( g = p(x) = \pi p(y) = \pi f \). Conversely, for a particular process, \( X \), we can lift the distribution function \( g \) to the associated adjoint distribution and get \( \pi p \). Since \( Y \) is i.i.d. over the adjoint sample space one can measure information production simply by using Shannon entropy with the adequate Boltzmann factor \( k_Y \) (adapted to the distribution function \( f \)). The commutative diagram therefore defines the process specific generalised entropy \( S_X \) over the sample space of the process class \( \Phi^* = \pi^{-1} \Phi^* \).
For the proof we use the *minimal description length* (MDL) (see also Text 2 in S1 File), the length of the shortest encoding that fully represents the data. In this context, we shall see that information production is tightly related to the notion of Kolmogorov complexity [25–27]; For a brief discussion, see Text 4 in S1 File. We explicitly demonstrate the method in an example for the class of sample space reducing (SSR) processes [21] that provide simple, analytically tractable models for driven dissipative systems, that typically exhibit distribution functions that are power laws. Their generalised entropy is exactly known for arbitrary driving rates [6, 21].

The purpose of the paper is to show that indeed, $S_X(X) = k_Y H(Y)$, represents generalized entropies in the second way. The proof is given constructively in the following section.

### Results

The key tool used in the following are simple substitution rules, *parsing rules*, that allow us to reversibly re-code (possibly correlated) data streams into new symbol streams that no-longer carry structure; see Text 3 in S1 File. The structure of a parsing rule we refer to as a *template*. A particular substitution rule derived from a template we refer to as a *parsing rule*.

The simplest parsing rule template, that we refer to as the *elementary* template, can be denoted by $[r s \rightarrow m]$, meaning that two symbols $r$ and $s$ that appear together are substituted with a new symbol $m$. In the following we will associate $m$ also with the symbol index, i.e. it is the $m$-th symbol in an alphabet.

To recode data one may choose a suitable set of parsing rule templates. We speak of a *relevant* set of parsing rule templates if (i) one can extract the full information content of a process, $X$, asymptotically, solely by using parsing rules from the set of templates, and (ii) if omitting one template from the set does not allow one to do so. In the following we focus on processes for which the elementary parsing rule template forms a relevant set. However, the arguments presented here extend naturally to more general sets of parsing rule templates; see Text 3 in S1 File.

A generalization of the framework we present from categorial to continuous random variables can be considered via first coarse-graining the continuous dynamics and then taking the limit to ever finer scales of coarse graining, however at considerable measure theoretic cost. Genuinely reformulating the theory in a framework of path-integrals will mainly be limited by translating the concepts of alphabet and parsing rules, which are inherently discrete objects, into a framework of continuous variables.

### Constructing adjoint process spaces

Decidability issues may forever limit our ability to design algorithms for reconstructing optimal adjoint representations from data that are not context specific. On the other hand, for any finite data over finite sample spaces and finite relevant set of parsing rule templates one can in principle find the optimal proxy to an adjoint representation by extensive search algorithms, even though time issues forbid such extensive searches in practical implementations. The general procedure underlying the construction of adjoint representations always remains the same.

Suppose $X$ is a process that emits symbols $i$ drawn from the alphabet $A_0 = \{1, 2, \ldots, W\}$, where $W$ is the number of symbols. $X$ generates data streams, $x(t) = x_t x_{t-1} \cdots x_1$, where every $x_t$ is one of the available symbols in $A_0$, which contains $W$ elements. Consider now two letters $r_1$ and $s_1$ such that the pair $r_1 s_1 \equiv x_{\tau} x_{\tau-1}$ for some positions $\tau$ in the data, has been identified to contain relevant information (e.g. because the pair is over-expressed), then we can rewrite the pair $r_1 s_1$ by a new letter $m_1 = W + 1$, which will become the first letter extending alphabet $A_0$ to $A_1 = A_0 \cup \{W + 1\}$, with the parsing rule $\pi_1 = [r_1 s_1 \rightarrow W + 1]$. 
We can iterate and produce parsing rules \( \pi_n = [r_n, s_n \rightarrow W + n] \), with letter indices \( r_n < W + n \) and \( s_n < W + n \). Where \( \pi_n \) maps data over the alphabet \( \mathcal{A}_{n-1} \) to data over \( \mathcal{A}_n = \mathcal{A}_{n-1} \cup \{W + n\} \). Note that the parsing rules \( \pi_n \) can be uniquely inverted, i.e. we can expand data over \( \mathcal{A}_n \) to data over \( \mathcal{A}_{n-1} \) using the inverse map \( \pi_n^{-1} = [W + n \rightarrow r_n, s_n] \). In other words the inverse parsing rules can be thought of being part of a “generative grammar”, [23]. We therefore can construct a sequence of maps \( \pi(n) = \pi_n\pi_{n-1} \ldots \pi_1 \) such that data \( x \) can be mapped to representations \( y_n = \pi(n)x \). At every parsing level, \( n \), we get a corresponding distribution function of the re-coded data, \( p_y(y_n) \), with a letter index \( z = 1, 2, \ldots, W + n \).

The Kraft and McMillan theorems [28, 29] tells us that if all that we know about a process are its marginal relative frequencies, \( g_i \), at which symbols \( i \) occur, there exists a shortest reversibly encoding of the data, \( x \), of characteristic length, \( L^\text{min} \), the minimal description length (MDL) that gives the theoretically achievable minimal length of \( x \) (in units of bits). \( L^\text{min} \) is a lower bound for the true MDL, \( L(x) \), that can only be attained asymptotically. The theorems state that

\[
kH(g) = \lim_{t \rightarrow \infty} \frac{1}{t} L(x(t)) ,
\]

with \( k = 1/\log(b) \), where \( b \) is the basis in which information is measured. For bits we typically have \( b = 2 \). \( kH(g) \) is the MDL in bits per symbol and \( L^\text{min} = tkH(g) \) is the minimal number of bits required to encode messages of length \( t \).

For data \( x(t) \) of length \( t \) we find a sequence of representations \( y_n(t) = \pi(n)x(t) \). Suppose now that for every \( t \) we find a parsing level \( n^*(t) \) such that \( y_{n^*(t)}(t) \) is a representation of the data \( x(t) \) that cannot be distinguished from an i.i.d. process (which we indicate here by a *). It follows that \( y_{n^*(t)}(t) \) is entirely determined by its marginal distribution of letters \( p_y(y_{n^*(t)}(t)) \) and obtain asymptotically \( kH(p_{y_{n^*(t)}(t)}) \approx \frac{1}{|y_{n^*(t)}(t)|} L^\text{min}(y_{n^*(t)}(t)) \), where \( \approx \) means asymptotically identical. With \( |.| \) we indicate the length of the sequence of letters in numbers of letters of the underlying alphabet. For instance we have \( |x(t)| = t \) and \( |y_{n^*(t)}(t)| \leq |y_n(t)| \). Then, we can asymptotically measure the information production of the process \( X \) to be

\[
\frac{1}{t} L^\text{min}(x(t)) \approx \frac{|y_{n^*(t)}(t)|}{|x(t)|} kH(p_{y_{n^*(t)}(t)}) .
\]

As discussed above, the “complexity” of a symbol \( \ell(z) \) is the number of letters in the original alphabet it codes for. As a consequence we can compute the average symbol complexity for data \( y_{n^*(t)}(t) \) to be given by \( \langle \ell \rangle_{y_{n^*(t)}} \approx \frac{|x(t)|}{|y_{n^*(t)}(t)|} \). As a consequence, we get for the adjoint process \( Y = \lim_t \rightarrow \infty Y_{n^*(t)} = \lim_t \rightarrow \infty \ell_{y_{n^*(t)}} \). The adequate unit of information, \( k_Y \), for measuring information production, \( k_Y H(p(Y)) \), therefore is given by

\[
k_Y(x) = \frac{k}{\langle \ell \rangle_Y} .
\]

For simplicity, suppose there is a maximal \( n^* \) that holds for all \( t \). Since we assumed that \( y_{n^*} \) is already indistinguishable from an i.i.d. process, applying another parsing rule would only compress the data without changing its MDL. This means, \( k_{y_{n^*(t)}} H(p(Y_{n^*(t)})) = k_{y^*} H(p(Y_{n^*})) \).

If, conversely, we look at a parsing level, \( n \), where the adjoint process is not yet i.i.d., then we can find a parsing rule, \( \pi_{n+1} \), such that \( k_{y_{n+1}} H(p(Y_{n+1})) < k_{y^*} H(p(Y_{n^*})) \). Additional knowledge always reduces the attainable information production rate.

In principle, for any finite amount of data \( x(t) \) one can construct the optimal map, \( \pi(n) \), for the process, \( X \), by minimizing over all possible sequences of parsing rules, at any fixed parsing
level $n$, if only we know the relevant set of parsing rules to consider and this set is finite. Then we can in principle find $n = n^*(t)$ such that no further reduction of the minimal description length is possible by applying any more parsing rules. In practice, an extensive search over all possible sequences of parsing rules is of course not feasible, even if the set of parsing rule templates only consists of the elementary template, and algorithms for inferring adjoint representations of data need to turn to different means of optimization. For theoretical considerations we may, however, assume that for a given finite relevant set of parsing rule templates and any finite $t$ we can find the optimal map $\pi$ (or one of several possible optimal maps if the map is not unique) or at least a map reasonably close to optimal, since the number of possible maps we would have to evaluate remains finite, too. Intuitively it is clear however, given an unknown process $X$ for which we cannot pre determine the respective relevant set of parsing rule templates, that one typically can no longer decide whether a map $\pi$ is optimal or not.

However, given an optimal $\pi$, the adjoint i.i.d. process, $Y$, is fully characterized by its marginal distribution, $f = p(Y)$, over symbols in the adjoint alphabet, $A^* = A_n$, and the information production, $I(X) = k Y H(f)$, is given by the Shannon entropy of $Y$. On the adjoint process space $\Phi^*$ we can use the measures of Shannon entropy, cross-entropy, and Kullback-Leibler information divergence, given that we use the appropriate unit of information, $k_Y$ of Eq (5). Since $Y$ is i.i.d. over $A^*$, the adjoint space naturally belongs to the family, $\Phi^*$, of all i.i.d. processes over this alphabet. Further, any $Y$ in $\Phi^*$ is fully characterized by its marginal distribution, $f'$, and the pair $(f', \pi)$, determines the process $X' = \pi^{-1} Y'$. Hence, the process class, $\Phi_\pi$, that naturally generalizes a process, $X$, with adjoint i.i.d. process, $Y = \pi X$, is given by $\Phi_\pi = \pi^{-1} \Phi^*$.

This construction completes the first part of the proof that establishes that we can essentially measure information production of a process as the Shannon entropy of the adjoint process. This entropy, however uses the marginal distributions over the adjoint alphabet as arguments and can therefore not be identified directly with the generalised entropies that use the marginal distributions over the original alphabet as arguments. In the next step we will pull the information measures over the adjoint message space back to the original message space.

### Information measures over extended alphabets

Suppose we have a process $X$ with an adjoint process $Y = \pi X$. For data, $x$, and its adjoint sequence, $y = y_n$, we obtain two histograms, $h_i(x)$, of symbols $i \in A_0$, and, $h_z(y)$, of symbols $z \in A_n$, respectively. The associated relative frequency distributions are given by $g = p(x) = h(x)/|x|$ and $f = p(y) = h(y)/|y|$. Further, every symbol, $z$, represents a number of $\ell(z)$ symbols, $\pi^{-1} z$, in the original alphabet with $\pi = \pi(n)$. We define $\delta_\pi(\pi^{-1} z, i)$ as the histogram of letters $i \in A_0$, that are parsed together into the symbol $z \in A_n$. For $z \in W$ where $W = |A_n|$, we have, $\delta_\pi(\pi^{-1} z, i) = \delta_i$ and $h_i(x) = \sum_{z \in A_n} \delta_\pi(\pi^{-1} z, i) h_z(y)$ needs to hold for all $i \in A_n$. This provides us with constraints,

$$ h_i(x) = \sum_{z \in A_n} \delta_\pi(\pi^{-1} z, i) h_z(y), \quad (6) $$

that we need in the next section. As a consequence, we have

$$ g_i(x) = \frac{|y|}{|x|} \sum_{z \in A_n} \delta_\pi(\pi^{-1} z, i) f_z(y). \quad (7) $$
We drop the arguments \((x)\) and \((y)\) (or \((y_n)\)) from now on and distinguish histograms by their index \(i\) (over \(A_i\)) and \(z\) (over the adoint alphabet \(A^* = A_n\)). Let \(\langle \ell \rangle_f = \sum_{z \in A^*} f(z)\ell(z)\) (note that in this notation we identify \(\ell\) \(\equiv \langle \ell \rangle_Y\) and \(\langle h_i \rangle_f = \sum_{z \in A} f(z)h_i(z)\) be the expectation values under the distribution \(f\), then, by construction, \(x = \sum_{z \in A} h_i(z), y = \sum_{z \in A} h_z\), and \(k_i(x) = k(\ell_j)^{-1}\). This means that we can write the constraints that link distributions \(g\) over \(A_n\), Eq (6), with distributions \(f\) over \(A^*\) in the following way

\[
0 = C_i(g|f) \equiv (\ell_j)^{-1}(\hat{h}_j)_f - g_i.
\]  

(8)

As mentioned before, the process class, \(\Phi_m\) that \(X\) belongs to, is completely determined by the map \(\pi\) and the process \(X\), by the pair \((f, \pi)\), see Fig 1. Therefore, we can identify the entropy of \(X\) with

\[
S_x(f) \equiv k_f H(f),
\]

with the process-specific Boltzmann factor, \(k_f \equiv k(\ell_j)\). For processes, \(X\) and \(X'\), and with \(f' = p(\pi X')\), the cross-entropy and the information divergence are

\[
S_{\pi}(f||f') \equiv k_f H_{\pi}(f||f')
\]

(10)

\[
D_x(f'||f) \equiv k_f D_{kl}(f'||f).
\]

In the special case where \(X\) is already an i.i.d. process, no features can be extracted from the data and \(n = 0, \pi = \pi(0) = id\), and \(\ell(z) = 1\), for all \(z = 1 \cdots W\). Consequently, \(S_x = kH, S_{\pi} = kH_{\pi}, D_x = kD_{kl}\) (Kullback-Leibler divergence), as required.

### Pulling back entropies to the original alphabet

In the next step one can construct entropy functionals over the original alphabet, \(A_n\), by lifting a distribution function, \(g'\), on \(A_n\), to a distribution function, \(f'\), over \(A^*\) by assuming that \(f = p(\pi X)\) is the true distribution function of the process, \(Y = \pi X\). We proceed by minimizing the information divergence, \(D_x(f'||f)\), with respect to \(f'\). More precisely, we minimize the functional \(\psi(f', \alpha, \eta)\) given by

\[
D_x(f'||f) - \alpha(|f'| - 1) - \sum_{i \in A_n} \eta_i C_i(g'|f')
\]

(11)

with Lagrange multipliers, \(\alpha\) and \(\eta_i\), that normalize \(f'\) and guarantee the constraints from Eq (8). Solving the variational principle \(\delta \psi = 0\) estimates \(f'\) at the minimum that is compatible with \(g'\). We identify this minimizer as \(\hat{f}(g')\) and obtain \(\alpha = 1/(\ell_j)\) and

\[
\hat{f}_{ij}(g') = \frac{D_v(\ell_j||f') + \sum_{i \in A_n} \eta_i (\hat{h}_j(z) - g_i^*)}{\ell_j} \ell_j(z),
\]

(12)

which has to be solved self consistently. If \(f\) already meets all matching constraints with \(g'\), i.e. if \(g' = g\) with \(g = p(X)\), then we have \(D_x(\hat{f}(g)||f) = 0\) and \(\hat{f}(g) = f\).

This means that one can lift marginal distributions, \(g'\), on \(A_n\) to distributions, \(\hat{f}(g')\), on \(A^*_n\) with respect to a particular process, \(X\). As a consequence one can pull the entropy, cross-entropy, and divergence back from distributions, \(f\), over the adoint sample space, \(A_n\), to distributions, \(g\), over the initial alphabet, \(A_n\), with respect to a particular process, \(X \in \Phi_n\). In
particular, one can define the projection operator, \( \pi \), through \( g = p(\pi^{-1}y) \equiv \pi p(y) = \pi f \) and the operator, \( \pi^*_\gamma \), that lifts distributions, \( g \), over alphabet, \( A_\gamma \), to distributions \( f \) over the extended alphabet, \( A^*_\gamma \), through

\[
\pi^*_\gamma g \equiv \minarg_{f' \mid g \otimes f'} D_n(f'\|p(\pi X)) ,
\]

with respect to the process \( X \), i.e., with respect to the distribution function, \( p(\pi X) \), of the i.i.d. process, \( \pi X \), over the adjoint alphabet. The lift operator gives us the minimizer \( f^*(g) = \pi^*_\gamma g \).

We find that \( \text{id} = \pi, \pi^*_\gamma \) and identify

\[
Su^\gamma(g) \equiv Su^\gamma(\pi^*_\gamma g)
\]

\[
S^{\text{cross}}u^\gamma(g'\|g) \equiv S^{\text{cross}}u^\gamma(\pi^*_\gamma g'\|\pi^*_\gamma g)
\]

\[
D_n(g'\|g) \equiv D_n(\pi^*_\gamma g'\|\pi^*_\gamma g),
\]

where, typically, \( g = p(X) \). We call those measures the pull-back entropy, cross-entropy, and information divergence of the process \( X \). Note, that while \( S_u, S^{\text{cross}}u, \) and \( D_n \) are universal on the entire class of processes \( \Phi_n \), pulling the information measures back to marginal distributions \( g \) over \( A_0 \) yields information measures that are specific to a particular process \( X \in \Phi_n \).

**Generalised entropies over initial alphabets**

The final question is how the pull-back measures \( S_u, S^{\text{cross}}u, \) and \( D_n \), defined in Eq (14), are related to generalised entropy functionals as derived for example in [6, 21]. There, functionals were derived to obtain the most likely histogram, \( h \), observed in a given process after \( t \) observations (for large \( t \)). Even for non i.i.d. processes, often the probability, \( P(h|\theta) \), to observe the particular histogram, \( h \), for \( t = \Sigma_{i=1}^n h_i \) observations factorizes \( P(h|\theta) = M(h) G(h|\theta) \) into a multiplicity, \( M(h) \), and a probability term of the sequences, \( G(h|\theta) \). \( \theta \) is a set of parameters that determines the process, \( X \) —it defines and parametrizes the process class \( X \) belongs to. Whenever such a factorization is possible, one can show that a generalised maximum entropy principle exists. Using the Boltzmann definition of entropy, the logarithm of multiplicity,

\[
\hat{S} = \log M/t, \text{ and defining } \hat{S}^{\text{cross}} = -\log G/t, \text{ and a generalised information divergence as } \hat{D}(g|\theta) = \frac{1}{t} \log P(h|\theta), \text{ where } g = h/t, \text{ the standard relations } \hat{D} = \hat{S}^{\text{cross}} - \hat{S} \text{ remain valid [5].}
\]

If we are looking at a family of processes \( X(\theta) \) rather than a single process \( X \) then we can no longer assume a priori that the same map \( \pi^{(1)} \) that takes some process \( X(\theta^{(1)}) \) to an adjoint representation efficiently is the same map \( \pi^{(2)} \) that takes some process \( X(\theta^{(2)}) \) from the family efficiently to an adjoint representation. However, there are ways in which we can think of constructing a single map \( \pi \) to some extended alphabet \( A^* \) such that all processes \( X(\theta) \) decorrelate under the same map \( \pi \). In the weakest version we can start with some extended alphabet \( A^* \) and the space of all i.i.d. processes \( \Phi^* \) over that alphabet and a map \( \pi \) and consider the process family \( \Phi_\pi = \pi^{-1} \Phi^* \), in which case the natural parameters of the processes \( X(\theta) \) are given by \( \theta = f \), where \( f \) are distribution functions over \( A^* \). For a discussion of stronger versions see Text 6 in S1 File.

However, if above assumption is valid, then \( X(\theta) \) forms a sub-class of processes in \( \Phi_\pi \) with distribution functions, \( g(\theta) = p(X(\theta)) \), and adjoint distribution functions, \( f(\theta) = p(\pi X(\theta)) \). Similarly we have \( \tilde{\ell}(\theta) = \sum_{f_y(\theta)} \ell(z), y' = \pi x' \) and \( t = |x'| = |y'| \tilde{\ell}(\theta) \). And in the next step we note
that the probability of observing a histogram $h'$ under process $X(\theta)$ is given by
\[
P(h'|\theta) = \sum_{h:h|\theta} P(x) = \sum_{h:h|\theta} (e^{-x})^P(x) = \sum_{h:h|\theta} \left(\frac{h'}{h}\right) \sum_{z \in A^*} f_\theta(z)^{h_z} \approx \max_{h':h'} \left(\frac{t}{h'}\right) \prod_{z \in A^*} f_\theta(z) ^ {h_z} = P(h'|f(\theta))
\]
with $h'$ being the respective maximizing argument $h^*$, the histogram over the extended space. The expression $h' = \hat{h}h^*$ are the constraints between the histograms on the original and adjoint alphabet. Since $y$ are sampled i.i.d, in the last line
\[
P(h'|f(\theta)) = \left(\frac{t}{h^*}\right) \prod_{z \in A^*} f_\theta(z) ^ {h_z}.
\]
is the multinomial distribution of histograms $h^*$. Also note that in the one but last line of Eq (15) $\simeq$ stands for asymptotically equivalent, meaning, that the relative error we make in $\log P(h'|\theta)$ by replacing $P(h'|\theta)$ with $P(h^*|f(\theta))$ vanishes in probability as $t \to \infty$. For more details see Text 7 in S1 File. We note, that this asymptotic equivalence is a consequence of processes that can be mapped into an adjoint representation becoming i.i.d. there and therefore implicitly can be thought of being ergodic on sufficiently large time scales.

Taking logs and multiplying both sides with $-k/t$ and setting $g' = h'/t$ and $f^* = k^*/|k^*|$, we then obtain
\[
D(g'|\theta) \simeq \min_{g' = h'/t} k_{t(\theta)} D_{KL}(f^*|f(\theta)) = \min_{g' = g'^*} D_{z}(f^*|f(\theta)) \approx D_{X(\theta)}(g'|g(\theta)),
\]
where $\simeq$ again means asymptotically identical for large $t$; from the second to the third line we used the definition of the lift operator from Eq (13) and $D_{X}$ from Eq (14). In other words, we have shown that in the limit $t \to \infty$ the generalised information divergence $D(g'|\theta)$ is identical to the pull-back divergence $D_{X(\theta)}(g'|g(\theta))$ as a functional. As a consequence we can use that $D_{X} = S_{X^{\text{cross}}} - S_{X}$ and identify $S_{X}$ with the generalised entropy and $S_{X^{\text{cross}}}$ with the generalised cross entropy. We see that the generalised entropy, $\hat{S}$, for the processes family, $X(\theta)$, is given by $\hat{S}(g) = S_{X(\theta)}(g)$.

We conclude that for all process classes (at least those that that decorrelate over a common map $\pi$) there exist notions of entropy, $S_{X}$, cross-entropy, $S_{X^{\text{cross}}}$, and divergence, $D_{X}$, that behave in the usual way, namely, $D_{X} = S_{X^{\text{cross}}} - S_{X}$. This means that for such processes (at least asymptotically) the probability to observe a particular histogram, $P \simeq \exp(-tD_{X})$, factorizes into a multiplicity term, $M \simeq \exp(tS_{X})$, associated with entropy, and a sequence probability term, $G \simeq \exp(-tS_{X^{\text{cross}}})$, associated with cross-entropy. In other words, we have shown that Boltzmann’s entropy, the logarithm of multiplicity, remains the correct estimator of information production. For complex systems the multiplicity will differ from a multinomial coefficient in arbitrarily complicated ways that might even depend explicitly on system parameters $\theta$, which
would mean a violation of SK1 axiom. In other words, the Boltzmann pull-back entropy functional typically will be more complicated than the Shannon entropy functional and can even violate SK1—yet they are the appropriate generalizations of entropy, in terms of information production. We also learned that beyond such parametric families of generalised entropies, i.e. beyond the pull-back measures, we find the standard notions of entropy, cross-entropy, and divergence present in the adjoint alphabet, where the only thing that is not universal about the information measures is the process-specific Boltzmann constant, $k_Y$, that needs to be used.

Example: SSR processes

We now demonstrate explicitly that the generalised entropy that—according to the previous section—is identified with the pull-back entropy, $S_X$, indeed does measure information production. We do that by considering slowly driven sample space reducing (SSR) processes, $X$, for which the generalised entropy functional is exactly known [21]. SSR processes are models of driven non-equilibrium systems. They are characterized by the fact that as the process unfolds the number of states accessible to the process reduces when no driving is present [30]. In its simplest form, the process relaxes to a ground state from which it has to be restarted. One can think of the process as a ball bouncing down a staircase with random jump sizes. The ball can only jump to steps lower than the last step it visited. Once it reaches the bottom of the staircase one lifts the ball to the top of the staircase (driving), and kicks it down the staircase again. The stairs represent (energy) states, the lowest being 1, the highest is $W$. The process exhibits path-dependence in the relaxation part, the current through the system breaks detailed balance between states. Such processes exhibit a Zipf law in their distribution function.

The micro-states, $x$, of the SSR process are sequences of states with elements $x_n \in \{1, 2, \ldots, W\} \equiv \Omega$. The transition probabilities between states $j \to i$ are

$$q(jj) = \Theta(j - i)q_i/Q_{j-1} + \delta_{j,1}q_i$$

(18)

where the first term on the right hand side describes the relaxing part of the SSR process (transitions only happen from higher $i$ to lower states, $j$, i.e., when $j < i$) with prior distribution $q_i$ and cumulative distribution, $Q_j = \sum_{i=1}^j q_i$, $\Theta$ is the Heavyside function. The second term captures the (slow) driving of the process. Slow here means that the system is only driven once the SSR process reaches its lowest position $i = 1$. SSR processes are Markovian since transition probabilities depend only on the current position and it is ergodic since after the relaxation process the system is reset to any state with probability, $q_i$.

To understand the statistics of the process we are interested in the distribution of visits to the individual states. We define the macro-state to be the histogram, $h_i$, of visits of $X$ to state, $i$. It is possible to compute $S_{\text{SSR}}(p) = \frac{1}{t} \log M(k)$, where the multiplicity, $M(k)$, is the number of different sequences, $x$, of length $t$ with the same histogram $h$. One finds [21]

$$S_{\text{SSR}}(p) = - \sum_{i=2}^W p_i \log \frac{p_i}{p_1} + (p_1 - p_i) \log \frac{p_1 - p_i}{p_1},$$

(19)

where $p_i = h_i/t$ are the relative frequencies of observing a state $i$. Note that this is the Boltzmann entropy of the system, yet it is not of Shannon form since it is derived from a Markov, and not an independent sampling process. Similarly, one finds the cross-entropy

$$S_{\text{SSR}}^{\text{cross}}(p|q) = - \sum_{i=1}^W p_i \log q_i + \sum_{i=2}^W p_i \log Q_{i-1},$$

(20)
and by maximizing \( \psi = S_{\text{SSR}} - S_{\text{SSR}}^{\text{new}} \), (negative information divergence \( D \)), one obtains the characteristic Zipf distribution of the slowly driven SSR process [30]

\[
q_i = p_i \frac{q_i}{Q_i}.
\]

(21)

For the special case of \( q_i = 1/W \) for all \( i \), the Zipf distribution, \( p_i \sim i^{-\alpha} \), is obvious, since \( Q_i = i/W \) and \( q_i/Q_i = 1/i \). It continues to hold for "well-behaved" \( q_i \) [31]. For instance, if \( q_i \propto i^{-\alpha} \) for \( \alpha > -1 \), then \( Q_i \propto i^{1+\alpha} \) and again \( q_i/Q_i \propto 1/i \).

In the next step we will use a simple example of a slowly driven SSR process in order to demonstrate how using extended alphabets works and how the respective generalised entropy functional is the adequate measure of information production.

**Example of a small SSR system.** To demonstrate how a minimal adjoint alphabet for a slowly driven SSR process looks like, consider such a process over an initial alphabet of \( W = 4 \) symbols (numbers) representing the four states, \( i \in \mathbb{A}_0 = \{1, 2, 3, 4\} \). A SSR sequence in that alphabet might look like \( x = 42121432143121214 \). Remember that the \( q_i \) are normalized weights such that the probability to sample the state \( j < i \) conditional on the process being in state \( i \) is given by \( q_j/Q_i \), with \( Q_i = \sum_{j < i} q_j \) and by \( q_i \) if the system is in the ground state \( i = 1 \) and the system gets driven. One can think of the adjoint SSR alphabet, \( \mathbb{A}_\alpha \), as the union of \( \mathbb{A}_0 \) with the set of new symbols that represent all possible strictly monotonic decreasing sequences on \( \mathbb{A}_0 \), i.e., \( \mathbb{A}_\alpha = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\} \), where the new symbol "5" represents the sequence 21, "6" stands for 31, "7" for 321, "8" for 41, "9" for 421, "10" for 431, and "11" for 4321. Since we have 7 new symbols, \( n^* = 7 \) extending the alphabet of original symbols \( \mathbb{A}_0 \) we have a total of 11 symbols in the extended alphabet \( \mathbb{A}_\alpha \). The 7 parsing rules producing the new symbols are given by

\[
\begin{align*}
\pi_1 &= [2 \ 1 \rightarrow 5] \\
\pi_2 &= [3 \ 1 \rightarrow 6] \\
\pi_3 &= [3 \ 5 \rightarrow 7] \\
\pi_4 &= [4 \ 1 \rightarrow 8] \\
\pi_5 &= [4 \ 5 \rightarrow 9] \\
\pi_6 &= [4 \ 6 \rightarrow 10] \\
\pi_7 &= [4 \ 7 \rightarrow 11]
\end{align*}
\]

(22)

and the map \( \pi = \pi_0 \pi_1 \pi_2 \pi_3 \pi_4 \pi_5 \pi_6 \pi_7 \), which maps between messages written in the initial and the adjoint alphabet, can be constructed. We therefore can rewrite our example \( x = 42121432143121214 \) into \( \pi(1)x = 4554354315541 \cdots \), then \( \pi(2)x = 455435465541 \cdots \), \( \pi(3)x = 455437465541 \cdots \), \( \pi(4)x = 45543746565541 \cdots \), \( \pi(5)x = 954746558 \cdots \), \( \pi(6)x = 9547[10]558 \cdots \), and finally \( \pi(7)x = 95 [11][10]558 \cdots \). We now project a distribution function, \( f_i \) on \( \mathbb{A}_\alpha \) to a distribution function, \( g_i \) on \( \mathbb{A}_0 \). We remember that the letters 5 to 11 code for the following subsequence: \( \pi^{-1}5 = 21, \pi^{-1}6 = 31, \pi^{-1}7 = 321, \pi^{-1}8 = 41, \pi^{-1}9 = 421, \pi^{-1}10 = 431, \) and \( \pi^{-1}11 = 4321 \), and see that all the new letters with index 5 to 11 represent sequences-junks that contain a 1. That is \( \tilde{h}_i(z) = \delta_{i5} + \sum_{j=5}^{11} \delta_{i j} \). Letter 2 is part of the sequences-junks represented by the extended letters 5, 7, and 9. That is \( \tilde{h}_2 = \delta_{15} + \delta_{25} + \delta_{27} + \delta_{29} + \delta_{115} \). Similarly we can find \( \tilde{h}_i(z) \) for \( i = 3 \) and \( i = 4 \). As a consequence the distribution functions \( g_i \) of the message \( x \) in original letters \( i = 1, \cdots, 4 \) and the distribution function \( f_2 \) of the adjoint message \( \pi(7)x \) in
Information production and generalised entropies

extended letters \( z = 1, 2, \cdots, 11 \) are given by four equations:

\[
\begin{align*}
g_1 &= \frac{1}{Z} (f_1 + f_5 + f_6 + f_7 + \cdots + f_8 + f_9 + f_{10} + f_{11}) \\
g_2 &= \frac{1}{Z} (f_2 + f_5 + f_6 + f_7 + f_{11}) \\
g_3 &= \frac{1}{Z} (f_3 + f_6 + f_7 + f_{10} + f_{11}) \\
g_4 &= \frac{1}{Z} (f_4 + f_6 + f_7 + f_{10} + f_{11})
\end{align*}
\]

(23)

where \( Z \) is a normalization constant such that \( 1 = \sum_{i=1}^{11} g_i \). Note that after applying \( \pi \) to a SSR process yields (asymptotically) that \( f_2 = f_5 = f_4 = 0 \). We can now express the asymptotic relative frequencies, i.e. the probabilities, \( f_6 \) in terms of the weights \( q_i \) on the SSR states \( i = 1, 2, 3, 4 \), and get, \( f_5 = q_2, f_6 = q_4 q_1/(q_1 + q_2), f_7 = q_4 q_2/(q_1 + q_2), f_8 = q_4 q_1/(q_1 + q_2 + q_3) \), and so forth.

Inserting the expressions for \( f_{12} \) in Eq (23) one self-consistently obtains the marginal distribution on the original alphabet

\[
g_i = \frac{1}{Z} \frac{q_i}{Q_i},
\]

(24)

as predicted from Eq (21);--note that if \( q_1 = 1/W \) is uniform, then the solution \( q_i/Q_i = 1/i \) is exactly reproducing Zipf's law and for a broad variety of choices for \( q_i \) one obtains approximate Zipf laws. That means that \( f \) fulfills the matching constraints of Eq (8) exactly and therefore also the lift, \( \pi_i^e \), of the asymptotic marginal distribution function, \( g \), to \( f \) is exact and is given by \( f = \pi_i^e g \). That is, we can see in this simple example how the distribution function \( g \) over the original alphabet can be predicted from knowing the distribution function \( f \) of letters of the extended alphabet.

Since slowly driven SSR processes are in fact also Markov processes, meaning, they are processes where the probability to sample the state of the process at \( t + 1 \) only depends on the state the system at the previous time-step \( t \) (compare transition probabilities Eq (18), one can also proof that the respective generalised entropies are actually the adequate information measures in this context. It is well known that the information production of a Markov process can be measured by the so-called conditional entropy, \( S_{\text{cond}} \). This is a functional that depends on the probabilities \( p^{(2)} = p_{ij} \), that a symbol \( j \) follows symbol \( i \) in the process. The SSR entropy on the other hand depends on \( p^{(1)} = p_i \) is the marginal distribution of the symbols, \( i \). If \( p^{(2)} \) is the maximizer of the conditional entropy, or more precisely, the minimizer of the conditional information divergence, and \( p^{(1)} \) is the minimizer of the SSR information divergence, then both estimators of entropy, the conditional entropy and the SSR entropy, are identical, \( S_{\text{cond}}(p^{(2)}) \equiv S_{\text{SSR}}(p^{(1)}) \), for all choices of the system parameters \( q \). For details of the computation, see Text 5 in S1 File.

Discussion

We showed that by identifying the entropy of a process with its information production it is possible to consistently extend the fundamental notion of entropy in statistical physics –Boltzmann entropy– to non-i.i.d. processes and processes that operate out of equilibrium. This is done by identifying isomorphisms that map entire process classes to adjoint representations
where processes are i.i.d. The sample space (or alphabet) of the adjoint process is typically much larger than the sample space of the original process. The isomorphisms can be thought of concatenations of parsing rules that map strongly correlated segments in the original process to new symbols. Information production of the adjoint i.i.d. process is quantified by Shannon entropy. Pulling back the entropy measure in the adjoint space to the original sample space and comparing the resulting functional with the Boltzmann entropy (process-specific log multiplicity) establishes the asymptotic equivalence of the notion of generalised entropy and information production.

This provides a comprehensive image that consistently links information theory and the statistical physics of categorial non i.i.d. processes in a context-sensitive way that allows us to consistently associate a notion of entropy, a cross-entropy (representing the constraints of the maximum entropy principle), and an information divergence (or relative entropy) to complex processes. Context-sensitive means that the functional form of the entropy depends on the class of processes considered; the concept of entropy itself, information production from the information theoretic perspective and Boltzmann entropy from the physics perspective, remains untouched.

If an adjoint representation of one process is found, one can find the adjoint representations of an entire class of processes that all de-correlate in their representations over the same adjoint sample space. This means that there exists a natural way how processes implicitly define their own generalization to an entire process class. This is possible because the property of de-correlating over the same adjoint sample space implements an equivalence relation. This has important consequences since these equivalence classes of processes generalize the idea of an ensemble to non-i.i.d. processes which provides a concise way –grounded in first principles– to extend the program of statistical physics to complex processes and their macro variables.

Supporting information

S1 File. The supporting informations supply seven texts that cover: SI Text 1 Shannon’s example of random texts from different alphabets: Here Shannon’s examples from his seminal paper about information theory are given in order to give an intuitive demonstration on the effects of extending alphabets e.g. from letters to words. SI Text 2 Minimal description length, i.i.d. processes, and compression A brief discussion on how compression works for i.i.d. processes. SI Text 3 Generative grammars, parsing rules and parsing rule templates: A brief discussion of the role parsing rules and their role in Generative grammars. We also discuss more broadly how we distinguish a parsing rule template from a particular parsing rule. SI Text 4 Information production and Kolmogorov complexity: Some remarks on how Information production relates to Kolmogorov complexity. [SI Text 5 Detailed algebraic steps for Eq. (29): The algebraic steps leading up to Eq (29) of this paper are given in detail. SI Text 6 About conjugate representations of process families: A brief discussion of issues concerning the existence of adjoint representation of entire process families rather than adjoint representations of a single process. SI Text 7 Measure concentration, typicality and asymptotic equivalence: A more detailed discussion of what asymptotic equivalence means in Eq. (15), explaining in which sense the generalized information measures based on Boltzmann entropy are equivalent to the pull-back information measures derived in this paper. (PDF)

Author Contributions

Conceptualization: Rudolf Hanel.
Formal analysis: Rudolf Hanel.
Funding acquisition: Stefan Thurner.
Methodology: Rudolf Hanel, Stefan Thurner.
Writing – original draft: Rudolf Hanel.
Writing – review & editing: Rudolf Hanel, Stefan Thurner.

References

1. Shannon CE. 1948 A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27 379–423 & 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

2. Kaniadakis G, Lissia M, and Scarfone AM. 2005 Two-parameter deformations of logarithm, exponential, and entropy: A consistent framework for generalized statistical mechanics. *Phys. Rev. E* 71, 046128. PMID: 15903747

3. Tsallis C, Gell-Mann M, and Sato Y. 2005 Asymptotically scale-invariant occupancy of phase space makes the entropy Sq extensive. *Proc. Nat. Acad. Sci. USA* 102 15377–15382. https://doi.org/10.1073/pnas.0503807102 PMID: 16230624

4. Naudts J. 1974 A generalised entropy function. *Comm. Math. Phys.* 37 175–182. https://doi.org/10.1007/BF01646343

5. Hanel R, Thurner S, and Gell-Mann M. 2014 How multiplicity of random processes determines entropy: derivation of the maximum entropy principle for complex systems. *Proc. Nat. Acad. Sci. USA* 111 6905–6910. https://doi.org/10.1073/pnas.1406071111 PMID: 24782541

6. Hanel R, and Thurner S. 2018 Maximum configuration principle for driven systems with arbitrary driving. *Entropy* 20 838. https://doi.org/10.3390/e20110838 PMID: 33266562

7. Plastino A R and Plastino A. 1993 Stellar polytropes and Tsallis’ entropy. *Phys. Letters A* 174 384–386. https://doi.org/10.1016/0375-9601 (93)90195-6

8. Wald R. 1993 Black hole entropy is the Noether charge. *Phys. Rev. D* 48 R3427(R). https://doi.org/10.1103/PhysRevD.48.R3427 PMID: 10016675

9. Wen-Sheng Xu and Freed KF. 2015 Generalized Entropy Theory of Glass Formation in Polymer Melts with Specific Interactions. *Macromolecules* 48 2333–2343. https://doi.org/10.1021/acs.macromol.5b00144

10. Tsallis C, Mendes RS, and Plastino AR. 1998 The role of constraints within generalized nonextensive statistics. *Physica A* 261 534–554. https://doi.org/10.1016/S0378-4371(98)00437-3

11. Chavanis P-H. 2006 Coarse-grained distributions and superstatistics. *Physica A* 359 177–212 https://doi.org/10.1016/j.physa.2005.06.043

12. Byrnes CI, Georgiou TT, and Lindquist A. 2001 A Generalized Entropy Criterion for Nevanlinna-Pick Interpolation with Degree Constraint. *IEEE Trans. Automatic Control* 46 822–839. https://doi.org/10.1109/9.1109928584

13. Dudowicz J, Freed KF, and Douglas JF. 2014 Generalized entropy theory of glass formation. *J. Chem. Phys.* 141 234903.

14. Goldstein S and Lebowitz JL. 2004 On the (Boltzmann) entropy of non-equilibrium systems. *Physica D* 193 53–66. https://doi.org/10.1016/j.physd.2004.01.008

15. Lieb E and Yngvason J. 2013 The entropy concept for non-equilibrium states. *Proc. R. Soc. A* 469 20130408. https://doi.org/10.1098/rspa.2013.0408 PMID: 24101892

16. Thurner S, Corominas-Murtra B, and Hanel R. 2017 The three faces of entropy for complex systems: information, thermodynamics, and the maxent principle. *Phys. Rev. E* 96 032124.

17. Gao X, Gallicchio E, Roitberg AE. 2019 The generalized Boltzmann distribution is the only distribution in which the Gibbs-Shannon entropy equals the thermodynamic entropy. *J. Chem. Phys.* 151, 034113 https://doi.org/10.1063/1.5111933 PMID: 31325924

18. Lemenev A, Appert-Rolland C, and van Wijland F. 2007 Thermodynamic formalism for systems with Markov dynamics. *J. Stat. Phys* 127 51–106 https://doi.org/10.1007/s10955-006-9254-0

19. Pregowska A, Kaplan E, and Szczepanski J. 2019 How Far can Neural Correlations Reduce Uncertainty? Comparison of Information Transmission Rates for Markov and Bernoulli Processes. *Internat. J. of Neur. Syst.*, 29 1950003 https://doi.org/10.1142/S0129065719500035 PMID: 30841769
20. Hanel R, Thurner S, and Tsallis C. 2009 Limit distributions of scale-invariant probabilistic models of correlated random variables with the q-Gaussian as an explicit example. *Europ. Phys. J. B* 72 263–268. https://doi.org/10.1140/epjb/e2009-00330-1

21. Hanel R, Corominas-Murtra B, and Thurner S. 2017 Understanding frequency distributions of path-dependent processes with non-multinomial maximum entropy approaches. *New J. Phys* 19 033008. https://doi.org/10.1088/1367-2630/aa611d

22. Korbel J, Lindner SD, Hanel R, and Thurner S. 2021 Thermodynamics of structure-forming systems. *Nature Com.* 12 1127. https://doi.org/10.1038/s41467-021-21272-7 PMID: 33602947

23. Chomsky N. 1956 Three models for the description of language. *IRE Trans. Info. Theory* 2 113–124. https://doi.org/10.1109/TIT.1956.1056813

24. Chomsky N and Schützenberger MP. 1963. *The algebraic theory of context free languages*, in Braffort P, Hirschberg D (eds.) Computer Programming and Formal Systems. Amsterdam: North Holland. 118–161.

25. Chaitin GJ. 1974 Information-theoretic limitations of formal systems. *J. of the ACM* 21 403–434. https://doi.org/10.1145/321832.321839

26. Kolmogorov AN. 1963 On Tables of Random Numbers. reprint in. *Sankhya: The Indian Journal of Statistics* A 25 369–376. Kolmogorov A N. 1965 Three Approaches to the Quantitative Definition of Information. *Problems Inform. Trans.* 1 1–7.

27. Solomonoff R. 1964. A Formal Theory of Inductive Inference Part I & II. *Information and Control* 7 1–22 & 224–254. https://doi.org/10.1016/S0019-9958(64)90223-2

28. Kraft LG. 1949. *A device for quantizing, grouping, and coding amplitude modulated pulses* Cambridge, MA: MS Thesis, Massachusetts Institute of Technology, hdl:1721.1/12390.

29. McMillan B. 1956 Two inequalities implied by unique decipherability. *IEEE Trans. Info. Theory* 2 115–116. https://doi.org/10.1109/TIT.1956.1056818

30. Corominas-Murtra B, Hanel R, and Thurner S. 2015 Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Nat. Acad. Sci. USA* 112 5348–5353. https://doi.org/10.1073/pnas.1420946112 PMID: 25870294

31. Corominas-Murtra B, Hanel R, and Thurner S. 2016 Extreme robustness of scaling in sample space reducing processes explains Zipf-law in diffusion on directed network. *New J. of Phys.* 18 093010. https://doi.org/10.1088/1367-2630/18/9/093010