Three faces of entropy for complex systems: Information, thermodynamics, and the maximum entropy principle

Stefan Thurner,[1,2,3,4] Bernat Corominas-Murtra,[1,4] and Rudolf Hanel[1,4]

[1]*Section for the Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria*

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

[3]*IIASA, Schlossplatz 1, 2361 Laxenburg, Austria*

[4]*Complexity Science Hub Vienna, Josefstädterstrasse 39, A-1090 Vienna, Austria*

(Received 6 May 2017; revised manuscript received 4 August 2017; published 15 September 2017)

There are at least three distinct ways to conceptualize entropy: entropy as an extensive thermodynamic quantity of physical systems (Clausius, Boltzmann, Gibbs), entropy as a measure for information production of ergodic sources (Shannon), and entropy as a means for statistical inference on multinomial processes (Jaynes maximum entropy principle). Even though these notions represent fundamentally different concepts, the functional form of the entropy for thermodynamic systems in equilibrium, for ergodic sources in information theory, and for independent sampling processes in statistical systems, is degenerate, $H(p) = -\sum_i p_i \log p_i$. For many complex systems, which are typically history-dependent, nonergodic, and nonmultinomial, this is no longer the case. Here we show that for such processes, the three entropy concepts lead to different functional forms of entropy, which we will refer to as $S_{\rm EXT}$ for *extensive entropy*, $S_{\rm IT}$ for the *source information rate* in information theory, and $S_{\rm MEP}$ for the entropy functional that appears in the so-called maximum entropy principle, which characterizes the most likely observable distribution functions of a system. We explicitly compute these three entropy functionals for three concrete examples: for *Pólya urn processes*, which are simple self-reinforcing processes, for *sample-space-reducing* (SSR) processes, which are simple history dependent processes that are associated with power-law statistics, and finally for *multinomial mixture* processes.

DOI: 10.1103/PhysRevE.96.032124

I. INTRODUCTION

Historically, the notion of entropy emerged in conceptually distinct contexts. In physics, thermodynamic entropy S was introduced by Clausius as an extensive quantity that links temperature with heat [1,2]. Boltzmann could relate this thermodynamic entropy to the number of microstates W in a system,

$$S_{\rm B} = k_{\rm B} \ln W, \tag{1}$$

assuming that in equilibrium all microstates are equally probable [3]. A microstate is a particular configuration of the components of a system. For example, a microstate for the ideal gas describes the positions and momenta of the gas particles in the volume of interest. Suppose a system has W microstates. In the case of equilibrium, the probabilities p_i of sampling microstates $i \in (1, \ldots, W)$ do not depend on time. Random processes that are independently repeating the same random experiment are called *Bernoulli* processes.[1] They follow a *multinomial* statistics. Equilibrium processes therefore can be thought of as sequences of tossing a biased die with W faces, where each face i represents a microstate with weight p_i. In that case, the entropy functional reads

$$H(p) = -k_{\rm B} \sum_{i=1}^{W} p_i \log p_i, \tag{2}$$

which is often referred to as the Gibbs formula. We set $k_{\rm B} = 1$ in the following. For example, a system consisting of N independent spins has $W = 2^N$ microstates, and the probability to find a particular configuration i is $p_i = 1/W$. As a consequence, $H(p) \sim \log W = N \log 2$ scales extensively, i.e., it grows linearly with the number of degrees of freedom N. Obviously, for systems composed of independent components (or weak and local interactions), Eq. (2) scales *extensively*.

Since Boltzmann, we identify this *extensive* functional with thermodynamic entropy, $S_{\rm EXT} = H$.[2] This notion gave way to the success story of statistical mechanics.

Independently from physics, in the context of *information theory* (IT), a functionally identical notion of entropy appears [4–6]. In IT, H quantifies how efficiently a particular stream of information can be coded if the information source is an ergodic finite state machine with W states. The information production rate $S_{\rm IT} = H$ determines if information can be coded, transmitted through a noisy channel, and decoded in an error-free way.

In the attempt to formulate statistical physics in a way that is independent of the physics of particles or spins, the *maximum entropy principle* (MEP) was developed [7]. It is a way to address statistical inference problems that are not

[1]We also call processes Bernoulli processes if they have a finite number $W > 1$ of discrete states and not only 2, as is often assumed.

[2]We do not claim that $S_{\rm EXT}$ is the *thermodynamic* entropy of history-dependent processes in the same way as H corresponds to the thermodynamic entropy of equilibrium processes, which is given by H in its maximum configuration p^*. For $S_{\rm EXT}$ we merely *select* extensivity as a defining property. Since for history-dependent processes the entropy concepts are no longer degenerate in general, distinct entropy notions will characterize distinct interrelations between macrostate variables and describe different aspects of the actual thermodynamics of history-dependent processes.

confined to physics. Again the same functional H appears, $S_{\mathrm{MEP}} = H$. The MEP approach is explicitly grounded in the statistics of multinomial Bernoulli processes. It can be used to infer the so-called *maximum configuration* from particular data, i.e., the distribution p_i of states i that is the most likely to be observed and that dominates the overall behavior of a system.

What these three very different approaches have in common is that physical equilibrium processes, information production of ergodic sources, and multinomial statistics are all essentially Bernoulli processes. As we will show below, the particular functional form of H from Eq. (2) is a generic consequence of this fact. For this reason, the different entropy concepts appear degenerate in the sense that S_{EXT}, S_{IT}, and S_{MEP} all are expressed by the identical functional H, which for obvious reasons may be referred to as the Boltzmann-Gibbs-Shannon (BGS) entropy. Processes that are nonergodic, history-dependent, or that have long-term memory explicitly break this degeneracy, which demonstrates that $H(p)$ is by no means a universal functional that fits all purposes. We will show in detail that the three mentioned entropy concepts lead to distinct entropy functionals that have to be determined for every family of processes individually.

We introduce some notation. For physical systems one typically uses the configuration, for IT, the process picture. They are equivalent. We will use both. By X we denote a class of systems or of processes. A configuration in a physical system corresponds to a path that a process can take; paths are the microstates in the process picture. A class is parametrized by a set of parameters θ, which we write as $X(\theta)$. For example, the class X of Bernoulli processes is given by the prior probabilities q_i, and $\theta = (q_1, \ldots, q_W)$. *Sample space* is denoted by $\Omega = \{1,2, \ldots, W\}$. $W = W(X(\theta))$ is the number of distinct elements in Ω. Sequences $x(N) = (x_1, \ldots, x_N) \in \Omega^N$ are either *paths* sampled by a process of length N, $X(N,\theta) = (X_1, X_2, \ldots, X_N)$, or configurations of a system with N elements, $X(N,\theta)$ with $W(X(N,\theta)) = W^N$. We can distinguish W^N different paths that a process $X(N)$ can take, or W^N distinct configurations of a system $X(N)$.[3] The *histogram* of a sequence $k(x(N)) = (k_1, \ldots, k_W)$ keeps track of how often state i is visited in the sequence $x(N)$. $p_i = k_i/N$ are the relative frequencies and $p = (p_1, \ldots, p_W)$ is the distribution of relative frequencies. The *phase-space volume* is the number of configurations in which a system at a given resolution can be.

Section I A introduces the three entropy concepts in their respective contexts. Section II shows that entropies are degenerate for Bernoulli processes, and Sec. III deals with Pólya urn processes and derives their corresponding IT, thermodynamic (extensive), and MEP entropies. Sections IV and V do the same for sample-space-reducing processes and multinomial mixture models, respectively. Section VI concludes.

A. The three concepts of entropy

In the following, we discuss how the three notions of entropy arise in the contexts of information theory, thermodynamics, and the maximum entropy principle.

1. Information-theoretic entropy and entropy rate

Shannon's approach to information theory deals with the question of how many bits per letter are needed on average to transmit messages of a certain type through information channels and what happens if these channels are noisy. Consider an information source process $X(\theta)$. The sample space Ω in this context is called an *alphabet* of letters (or a lexicon of words) $i \in \{1, \ldots W\}$. $X(N,\theta)$ generates messages, i.e., realizations or samples $x(N) = (x_1, x_2, \ldots, x_N)$. To transmit the message through an information channel, one has to translate messages into a code that has b symbols in the code alphabet (typically the code is binary, $b = 2$) such that the average number of bits per letter becomes minimal. Intuitively this means that frequently observed letters are assigned short binary codewords while infrequent letters are assigned longer codewords.

Shannon identified four properties of a functional H—the four Shannon-Khinchin (SK) axioms—that measure the average amount of information (in bits) that is required to encode messages that generate letters i with probabilities p_i. Three of these properties are of a technical nature. SK axiom 1: H is a continuous function that depends on p only and no other variables; SK axiom 2: $H(p_1, \ldots, p_W)$ is maximal for the uniform distribution $p_i = 1/W$; SK axiom 3: $H(p_1, \ldots, p_W, 0) = H(p_1, \ldots, p_W)$. The fourth property, the so-called *composition axiom* (SK axiom 4), states that H measures information independent of the way the W states get sampled with the probabilities p_i. It states that a system composed of two systems A and B that are statistically dependent on the entropy of the composed system $S(AB) = S(A) + S(B|A)$ is the entropy of system A alone plus the entropy of system B, conditional on A. Details on conditional entropy follow below. SK axioms 1–4 determine H uniquely up to a multiplicative constant. H is the functional given in Eq. (2).

Two theorems, one by Kraft [5] and one by McMillan [6], assure us that there exists a practical family of uniquely decodable codes (the prefix codes) if and only if $\sum_{i\in\Omega} b^{-\ell_i} \leqslant 1$, where ℓ_i is the length of the codeword for letter i, and b is the size of the code alphabet. For a binary code $b = 2$ this means that if the source variables $X_n(\theta)$ are identically independently distributed (i.i.d.), or equivalently if letters i appear with fixed probabilities p_i for all n, one can find codewords of length ℓ such that $1 - \log_2(p_i) \geqslant \ell_i \geqslant -\log_2(p_i)$. Such a code requires the fewest bits for transmitting messages. Using $\log_2(p_i) = \log(p_i)/\log 2$ we have

$$1 + H(p)/\log 2 \geqslant \langle \ell \rangle \geqslant H(p)/\log 2. \tag{3}$$

This means that $H(p)$ establishes the lower bound for the so-called *information rate* or source information rate of i.i.d. processes in bits per letter for prefix codes.

[3] For Bernoulli processes this equivalence is trivial. It is equivalent to tossing N independent dice at once, or to tossing one die N times in a sequence. Both result in N i.i.d. random variables X_n, $n \in 1, \ldots, N$. The process and the system picture only differ in terms of how variables X_n may depend on other variables X_m. For processes there may exist a time ordering, where X_n depends on variables X_m that appeared earlier in time, $m < n$.

What if we are not encoding letters but entire parts of messages $x(N)$ that are sampled from Ω^N with respective probabilities $p(x(N))$? The information rate of $x(N)$ is generally defined as [8]

$$S_{\rm IT}(x(N)) = -\frac{1}{N}\log p(x_N, x_{N-1}, \ldots, x_2, x_1), \tag{4}$$

where the joint distribution appears. For processes in which each X_n may depend on earlier events, we can rewrite Eq. (4). Using the notions for the empty sequence $x(0) = \emptyset$ and for the initial distribution $p(i|\emptyset)$, we write $p(x(N)) = \prod_{n=1}^{N} p(x_n|x(n-1))$, and we obtain

$$S_{\rm IT}(x(N)) = -\sum_{n=1}^{N} \log p(x_n|x(n-1)). \tag{5}$$

The Shannon-McMillan-Breiman (SMB) theorem [4,9,10] states that for Markov chains with transition probabilities $p(i|j)$ and stationary distributions p_j, the asymptotic information rate is given by the *conditional entropy* $H(X_{n+1}|X_n)$, i.e.,

$$\lim_{N\to\infty} S_{\rm IT}(x(N)) = -\sum_{j=1}^{W} p_j \sum_{i=1}^{W} p(i|j)\log p(i|j). \tag{6}$$

For Bernoulli processes, where $p(i|j) = p_i$, obviously $\lim_{N\to\infty} S_{\rm IT}(x(N)) = H(p)$. Note that for history-dependent process classes X, the law of large numbers that plays a crucial role in the SMB theorem does not necessarily apply, and the situation needs to be analyzed carefully for each specific path-dependent process.

The SMB theorem states that for Markov chains one can transmit messages at lower bit rates, $H(X_{n+1}|X_n) \leqslant H(p)$, by using optimal code lengths $\ell_i(j) \sim -\log_2 p(i|j)$ that are conditioned on the most recent event j of a message, $\langle\ell\rangle \sim H(X_{n+1}|X_n)/\log 2$. Also history-dependent processes can in principle be coded more efficiently. However, this does not mean that the transmission of information becomes more efficient since the key (decoding table) to the constantly updated coding schemes must be transmitted in addition to the source information. The *effective information rate* measures the total amount of information the sender has to transmit to the receiver.

2. *Thermodynamics and extensive entropy*

Traditionally thermodynamics deals with "homogeneous" matter, such as ideal gases or solid bodies in thermal equilibrium, and it characterizes systems independent of size, shape, and scale in terms of so-called *intensive* variables, such as temperature and pressure. Conjugate variables, such as volume and entropy, relate the intensive variables to the number of system components, or more precisely to the number of degrees of freedom. If extensive variables do not scale linearly with the degrees of freedom, no reasonable thermodynamic equations will exist.

If two initially separated systems A and B (that are at the same temperature and pressure, with volumes V_A and V_B and thermodynamic entropies $S(A)$ and $S(B)$, respectively) are combined, this implies that $V_{AB} = V_A + V_B$ and $S(AB) = S(A) + S(B)$. The extensivity of the thermodynamic entropy results from particles being *indistinguishable*, meaning that permutations of indistinguishable particles do not change the microstate. This effectively resolves the Gibbs paradox by constraining particles to their independent share of the volume V/N; see, for example, [11].

Assume that $W = \bar{W}(X_n)$ is the number of states the nth particle can be in, say discrete positions in a container. Then, if N_A and N_B are the numbers of identical particles in the two containers, respectively, one finds that the *effective number of configurations* $\hat{W}$ in the combined container is given by $\hat{W}(AB) = W^{N_A+N_B} = W^{N_A}W^{N_A} = \hat{W}(A)\hat{W}(B)$. Boltzmann entropy $S_{\rm B} = \log\hat{W} = N\log W$ is extensive in N. When the states that each particle can be in are sampled from a given distribution q—which may not be uniform—one can still estimate the *effective* number of states as $\hat{W} \sim e^{NH(q)}$, where $H(q)$ is the Gibbs formula Eq. (2) for distribution q. As a consequence, $e^{H(q)}$ measures the effective amount of states per particle,[4] and Boltzmann entropy remains extensive, $\log\hat{W} = NH(q)$.

This is generally valid for systems or processes $X(N) = (X_1, \ldots, X_N)$ described by i.i.d. variables X_t. Systems or processes with strong constraints, strong interactions, with nonstationary prior probabilities for states $q_i(t)$, strong internal correlations, or with history-dependent dynamics, typically populate subspaces of the entire phase space, and H (Gibbs formula) is no longer extensive. For examples, see, e.g., [12–14]. In the more general case, one can estimate $\hat{W}(N)$ by

$$\hat{W}(N) \equiv \prod_{t=1}^{N} \bar{W}(X_t), \tag{7}$$

where, again, we measure $\bar{W}(X_t) \sim e^{H(q(t))}$. Such systems or processes are called nonextensive, and the SK axiom 4 (composition axiom) is violated. In this case, H lost the extensive property. However, one can find a functional expression for an entropy that remains extensive—even though the underlying system or process is nonextensive. We call such a functional the *extensive entropy*, $S_{\rm EXT}$. Since from Eq. (7) it follows that $\hat{W}(N)$ is monotonically increasing in N, an inverse function L_X exists such that $L_X(\hat{W}(N)) = N$, and a unique extensive trace-form functional can be found (see Appendix A),

$$S_{\rm EXT}(p) = \sum_{x\in\Omega^N} s(p(x)) = Ns_0. \tag{8}$$

Here $p(x)$ is the probability to sample path x, and s_0 is a constant.

For classes X, which are compatible with the first three SK axioms 1–3, but violate SK axiom 4 (often nonergodic processes), all extensive entropies $S_{\rm EXT}$ can be classified by

[4]Alternatively, one can measure the first moment of the rank $r(i|q)$ of states i with respect to the distribution function q. The rank $r(i|q)$ is a permutation on Ω, such that $r(i|q) > r(j|q)$ if $q_i > q_j$. For a reference process being concentrated uniformly on $\bar{W}$ states, one finds $\langle r\rangle_n \equiv \sum_{i=1}^{W} q_i(n)r(i|q(n)) = (\bar{W}+1)/2$. Conversely, one may define $\bar{W}(X_n) \equiv 2\langle r\rangle_n - 1$.

(c,d) entropies, [15]. These, in a convenient representation, take the form

$$S_{(c,d)}(p) = \frac{\frac{e}{c}\sum_{i=1}^{W}\Gamma(1+d,1-c\log(p_i))-1}{1-c+cd}. \quad (9)$$

$S_{(c,d)}$ is parametrized by two scaling exponents c and d that characterize the asymptotic scaling behavior of the entropy of the nonextensive system or process. The exponents are one-to-one related with the phase space of the system [12], and they can be computed using

$$\frac{1}{1-c} = \lim_{N\to\infty} N\frac{d}{dN}\log\hat{W}(N),$$
$$d = \lim_{N\to\infty}\log\hat{W}(N)\left(c-1+\frac{1}{N\frac{d}{dN}\log\hat{W}(N)}\right). \quad (10)$$

(c,d) entropies are extensive quantities for nonextensive system classes.

Extensive systems correspond to the special case $c=1$ and $d=1$, and one finds $\frac{1}{e}S_{(1,1)}(ep) = -\sum_i p_i\log p_i$ (e is the Euler constant). The special case of $d=0$ corresponds to power laws and recovers Tsallis entropy [16], $\frac{1}{\eta}S_{(c,0)}(\eta p) = (1-\sum_i p_i^c)/(c-1)$, where $\eta = c^{1/(c-1)}$ (note that $\lim_{c\to 1}\eta = e$). For $c<1$, (c,d) entropies describe the phase-space growth of so-called winner-take-all processes (WTA), where probabilities p_i of sampling states $i\in\Omega$ concentrate over time in one single element $j\in\Omega$, the winner, and $\lim_{n\to\infty}\bar{W}(X_n)=1$. WTA processes also violate SK axiom 3.

3. The entropy of the maximum entropy principle

The MEP is closely related to the question of finding the most likely observable macroscopic property (macrostate) of a system or a process. The distribution function p, or the histogram k, of events x_n that occurred along the path x of a process $X(N,\theta)$ is such a macrostate. In other words, how do we find the most likely distribution function of a given process or a system? Denoting the probability of finding the histogram by $P(k|\theta)$, the most likely histogram k^* is obtained by maximizing $P(k|\theta)$ with respect to k under the constraint $\sum_i k_i = N$. k^* is the best predictor for observing a macrostate that is generated by the process $X(N,\theta)$. If P becomes sharply peaked as N becomes large, predictions will become very accurate.

The MEP of the process $X(N,\theta)$ is obtained by factorizing P into two terms, $P(k|\theta) = M(k)G(k|\theta)$. M can sometimes be identified with the *multiplicity* of the macrostate k, i.e., the number of microstates that lead to the macrostate. This is certainly true for Bernoulli processes, see Sec. II C, and for SSR processes, see Sec. IV. For Bernoulli processes, $M(k)$ is equivalent to the multinomial factor $\binom{N}{k}$, while for SSR processes $M(k)$ is a different type of multiplicity factor. Similarly, $G(k|\theta)$ can sometimes be identified with the probability of a microstate belonging to k. In other cases (e.g., for Pólya urn processes, see Sec. III) such a factorization $P=MG$ exists, but neither M nor G has an immediate interpretation as a multiplicity or as the probability to observe a particular microstate. However, if such a factorization can be defined in a meaningful way, not only a *minimum relative entropy principle*, but also a corresponding *maximum entropy principle* exists.

Taking logarithms $\log P = \log M + \log G$ does not change the location $k^* = k$ of the maximum of $P(k|\theta)$, and

$$\underbrace{\frac{1}{f}\log P(k|\theta)}_{-S_{\rm rel}} = \underbrace{\frac{1}{f}\log M(k)}_{S_{\rm MEP}} + \underbrace{\frac{1}{f}\log G(k|\theta)}_{-S_{\rm cross}}. \quad (11)$$

Here f is an appropriate scaling factor, which corresponds to the degrees of freedom of microstates; see [13].

$S_{\rm rel}$ is the (generalized) *relative entropy* or *information divergence*. Note that for Bernoulli processes, where θ is given by the prior probabilities q, and $P(k|\theta)$ is the multinomial distribution function, $S_{\rm rel}$ is identical to the Kullback-Leibler divergence [17], $H_{\rm rel}(p|q) \equiv D_{\rm KL}(p||q) = \sum_i p_i(\log p_i - \log q_i)$.

$S_{\rm MEP} = \frac{1}{f}\log M(k)$ is the (generalized) entropy that appears in the MEP, which we call *MEP entropy*. It is sometimes called the *reduced* Boltzmann entropy,[5] which is defined as $s_B = S_B/f$. This name is justified whenever M can be interpreted as a multiplicity factor.

$S_{\rm cross}(p|\theta) = -\frac{1}{f}\log G(k|\theta)$ is the (generalized) *cross-entropy*, which represents sets of constraints imposed by the parametrization θ. Again, for Bernoulli processes with prior probabilities q, the cross entropy takes the well-known form

$$H_{\rm cross}(p|q) = -\sum_{i=1}^{W} p_i\log q_i. \quad (12)$$

Note that within a maximum configuration approach, not only the notion of entropy but also the notions of cross-entropy and relative entropy, i.e., information divergence, can be naturally generalized. For Bernoulli processes, these notions correspond to H, $H_{\rm rel}$, and $H_{\rm cross}$. Moreover, the relation $S_{\rm rel} = S_{\rm cross} - S_{\rm MEP}$ is also valid in the generalized form.

II. BERNOULLI PROCESSES

We compute the three entropies $S_{\rm IT}$, $S_{\rm EXT}$, and $S_{\rm MEP}$ for Bernoulli processes and show that they are identical with H from Eq. (2). Bernoulli processes have no memory, and states $i=1,\ldots,W$ are sampled independently from the prior probability distribution $q=(q_1,\ldots,q_W)$. Bernoulli processes of length N, $X(N,\theta)$ are parametrized by $\theta\equiv q$.

Consider the histograms k with $\sum_{i=1}^{W}k_i = N$ as the macrostates of the Bernoulli process, and sequences $x(N)$ as their microstates. Then the probability to sample a particular sequence $x(N)$ with histogram k is given by $G(k|q) = \prod_{i=1}^{W} q_i^{k_i}$. The multiplicity $M(k)$ is given by the multinomial factor $M(k) = \binom{N}{k}$, and the probability to sample histogram k is $P(k|q) = M(k)G(k|q)$. The number of degrees of freedom of a sequence of length N is $f=N$.

[5] Boltzmann's principle as formulated by Planck [18] identifies entropy $S_{\rm B}$ with the logarithm of multiplicity, $S_{\rm B} = k_{\rm B}\log M$.

A. The information rate of Bernoulli processes

Since Bernoulli processes have no memory, the transition probabilities $p(i|x,\theta) = q_i$ do not depend on path x. The information rate from Eq. (5) is

$$\begin{aligned} S_{\rm IT}(x) &= -\frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\, x(n-1),\theta) \\ &= -\frac{1}{N}\sum_{i=1}^{W} k_i \log q_i \\ &= -\sum_{i=1}^{W} p_i \log q_i = H_{\rm cross}(p|q). \end{aligned} \tag{13}$$

Since $\lim_{N\to\infty} p = q$, for a typical sequence x one finds

$$\lim_{N\to\infty} S_{\rm IT}(x) = H_{\rm cross}(p|q) = H(p). \tag{14}$$

The entropy $S_{\rm IT} = H(p)$ measures the typical information rate for optimally coded Bernoulli processes.

B. The extensive entropy of Bernoulli processes

There are three ways to see what the extensive entropy for the Bernoulli process $X(N,\theta)$ is.

(i) Since Bernoulli processes fulfill all four Shannon Khinchin axioms, a well-known theorem by Shannon [4] (Appendix B) states that $S_{\rm EXT}(p) = H(p)$.

(ii) The effective number of configurations $\hat{W}(N) = \bar{W}(X(N))$ of a Bernoulli process $X(N)$ grows exponentially, $\hat{W}(N) = \bar{W}^N$. This is because $X(N)$ is composed of N i.i.d. Bernoulli trials X_n. Using Eq. (A5) and setting $\hat{W}(N) = \bar{W}^N$, we see that $L_X(y) = \log y/\log \bar{W}$. As a consequence, $s_0 = \log \bar{W}$ and $s(y) = -y\log y$, meaning that $S_{\rm EXT}(p) = \sum_i s(p_i) = H(p)$.

(iii) Using Eq. (10) and the exponential phase-space growth, $\hat{W}(N) = \bar{W}^N$, of Bernoulli processes, one verifies that $c = 1$ and $d = 1$. To obtain c, one computes

$$\frac{1}{1-c} = \lim_{N\to\infty} N\frac{d}{dN} N\log \bar{W} = \infty. \tag{15}$$

As a consequence, $c = 1$. Similarly one obtains $d = 1$. Since $S_{(1,1)} = H$, we conclude that $S_{\rm EXT} = H$.

C. The MEP entropy of Bernoulli processes

Since for Bernoulli processes the degrees of freedom are simply given by the number of samples $f = N$, using Stirling's approximation $N! \sim N^N e^{-N}$ it is easy to see that

$$\begin{aligned} S_{\rm MEP} &= \frac{1}{N}\log\binom{N}{k}, \\ \text{(Stirling)} &\sim \frac{1}{N}\log\frac{N^N}{\prod_{i=1}^{W} k_i^{k_i}} = -\frac{1}{N}\log\prod_{i=1}^{W} p_i^{k_i} \\ &= -\sum_{i=1}^{W} p_i \log p_i = H(p). \end{aligned} \tag{16}$$

The maximum entropy $S_{\rm MEP}$ of Bernoulli processes is again equivalent to $H(p)$.

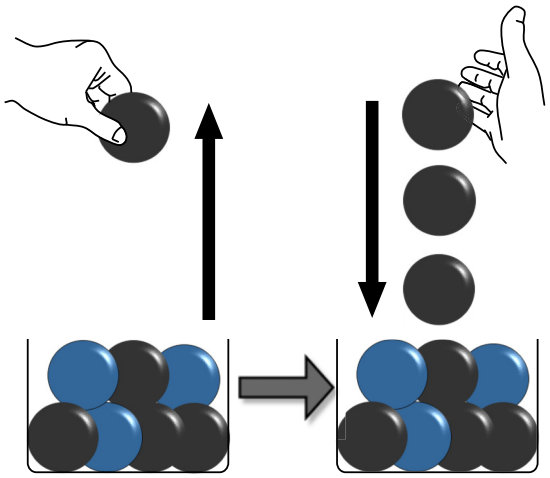

FIG. 1. Schematic illustration of a Pólya urn process. When a ball of a certain color is drawn, it is then replaced by $1+\delta$ balls of the same color (here $\delta = 2$). The process is repeated N times. This reinforcement process creates a history-dependent dynamics. After [19].

The relative entropy $S_{\rm rel} = -\frac{1}{f}\log P$ is given by the Kullback-Leibler divergence $D_{\rm KL}$,

$$S_{\rm rel}(p|\theta) = \sum_{i=1}^{W} p_i(\log p_i - \log q_i) \equiv D_{\rm KL}(p||q). \tag{17}$$

The cross-entropy $S_{\rm cross} = -\frac{1}{N}\log G$ is given by $-\sum_i p_i \log q_i$ and imposes a linear first moment constraint on p in the MEP. This can be seen by reparametrizing q_i by $\exp(-\alpha - \beta\epsilon_i)$, which yields $S_{\rm rel} \sim H(p) - \alpha\sum_{i=1}^{W} p_i - \beta\sum_{i=1}^{W} p_i\epsilon_i$. α and β play the role of Lagrangian multipliers in the maximization problem. For Bernoulli processes, the maximum configuration asymptotically predicts $p^* = q$.

III. THE THREE ENTROPIES OF PÓLYA URN PROCESSES

A. Pólya urn processes

Multistate Pólya urn processes [20,21] are an abstract representation of path-dependent, self-reinforcing processes with memory. A Pólya urn is initially filled with a_i balls of color $i = 1,\ldots,W$. One draws the first ball of color $x_1 = i$ with probability $q_i = a_i/A$, where $A = \sum_{i=1}^{W} a_i$ is the total number of balls initially in the urn. If we draw a ball of color i, we do not only replace it, as we would do in a Bernoulli process (drawing with replacement), but we add another δ balls of the same color and thus reinforce the probability to draw color i in subsequent trials; see Fig. 1. As a consequence, the probability to draw another state (color) i after N samples drawn is given by

$$p(i|k,\theta) = \frac{a_i + k_i\delta}{A + N\delta} = \frac{q_i + k_i\gamma}{1 + N\gamma}, \tag{18}$$

where $\gamma = \delta/A$ is the *reinforcement* parameter, $q_i = a_i/A$, and $\theta = (q,\gamma)$ is the set of parameters characterizing the process. If $\gamma = 0$, the Pólya urn process is just "drawing with replacement" and it is the same as a Bernoulli process. If $\gamma > 0$, the probability to draw color i in the $(N+1)$th

sample depends on the history of samples x in terms of the histograms k.

Pólya urn processes and nonlinear versions of it exhibit a crossover between dynamics that is asymptotically a Bernoulli processes (weak reinforcement) and dynamics that is referred to as "winner take all" (WTA) dynamics (strong reinforcement). For intermediate reinforcement strengths γ, the details of random events that happen early on will determine whether the system behaves one way or the other. How sequences $x = (x_1, \ldots, x_N)$ behave for large N depends on the samples x_n that are drawn at times n much smaller than N.

Pólya processes operate at the edge of Bernoulli and WTA dynamics. If we measure the histogram $k(N_1)$ of the process after N_1 steps, we may continue the process by thinking of starting a different Pólya urn with an initial condition $k(N_1)$. For this we consider the histogram $k' = k - k(N_1)$ of $N' = N - N_1$ samples and define $a_i(N_1) = a_i + k_i(N_1)\delta$ and $A(N_1) = A + \delta N_1$. It is easy to see that one again is looking at a Pólya urn process. However, the parameters have been modified from $\theta = (q,\gamma)$ to $\theta' = (q',\gamma')$, where

$$q' = \frac{q_i + \gamma k_i(N_1)}{1 + \gamma N_1}\gamma, \quad \gamma' = \frac{\gamma}{1 + N_1\gamma}. \tag{19}$$

As a consequence, the effective reinforcement $\gamma' < \gamma$, and $\gamma'(N_1) \to 0$ as $N_1 \to \infty$. The distribution q' gets modified by the history of the process $x(N_1)$, and the effective reinforcement parameter γ' decreases over time. Whether a particular realization of a process defined by θ enters the WTA dynamics therefore depends on whether the modified Pólya urn with parameters θ' enters WTA dynamics or not. This depends on which path $x(N_1)$ the urn process took within the first N_1 steps. If in those first steps one of the elements i acquires most of the weight, the process can enter WTA dynamics, meaning that i eventually gets sampled almost all of the time. Nonlinear Pólya processes, where the effective reinforcement decays more slowly as time progresses, almost certainly enter WTA dynamics. We can now discuss the three entropies of Pólya processes.

B. The information rate of Pólya urn processes

In WTA scenarios, the relative frequency p_j to observe the winner j approaches 1, meaning that p concentrates on the winner state j, which essentially becomes the only state that is sampled,

$$p_j(N) \sim 1 - \frac{1 - q_j}{1 + \gamma N}. \tag{20}$$

Without knowing the exact distribution of the "loser" states $i \neq j$, we assume that all those states have equal probabilities,

$$p_i(N) = \frac{1 - p_j(N)}{W - 1} = \frac{1 - q_j}{(W-1)(1 + \gamma N)}. \tag{21}$$

Following Eq. (20), the information rate of a WTA process can be estimated,

$$N S_{\mathrm{IT}}(x) = -\sum_{n=1}^{N} \log p(x_n | x(n-1)) \sim \frac{1 - q_i}{\gamma} \log N + \text{const.} \tag{22}$$

The information rate of a Pólya process in the WTA mode asymptotically approaches zero. The total information production, i.e., the number of bits required to encode the entire sequence, grows logarithmically, $N S_{\mathrm{IT}}(x(N)) \propto \log N$. If the Pólya process does not enter WTA dynamics, it behaves like a Bernoulli process sampling from the limit distribution $p(\infty) = \lim_{N\to\infty} p(N)$ with an information rate, $H(p(\infty))$.

C. The extensive entropy of Pólya urn processes

The effective number $\bar{W}$ of a typical sequence x of length N, with j the winner in the WTA process, can be estimated by inserting q from Eqs. (20) and (21) into $\bar{W}(n) \sim \exp(H(q(n)))$, or alternatively by using this q to compute the first-rank moment $\langle r \rangle$ and $\bar{W}(n) \sim 2\langle r \rangle - 1$. One uses Eq. (7) to compute

$$\hat{W}_{\text{Pólya}}(N) \propto (1 + \gamma N)^{\frac{1-q_j}{\gamma}}. \tag{23}$$

From Eq. (10) it follows that $1/(1-c) = (1-q_j)/\gamma$. Therefore, the (c,d) class of Pólya urn processes is

$$c = 1 - \frac{\gamma}{1 - q_j} \quad \text{and} \quad d = 0. \tag{24}$$

Note that c is negative for γ sufficiently large, which means that the SK axiom 3 is violated by Pólya processes in WTA dynamics. As a consequence, (10) might no longer hold, since it was derived under the assumption that SK axioms 1–3 do hold. However, one can still safely compute the extensive entropy using Eq. (A5) with Eq. (23) to find

$$S_{\mathrm{EXT}}(q) = \frac{1}{1 - q_j} \sum_{i=1}^{W} q_i \log_c q_i = S_{(c,0)}, \tag{25}$$

where $\log_c(x) = (x^{1-c} - 1)/(1-c)$ and $c = 1 - \gamma/(1-q_j)$. This is exactly the result that we get from Eq. (10).

D. The MEP of Pólya urn processes

For Pólya urn processes, the probability to observe a sequence x is the same as the probability to observe any other sequences x' with the same histogram k. Therefore, $P(k|\theta) = M(k)G(k|\theta)$ factorizes into the multiplicity $M(k)$, which is given by the multinomial factor and the sequence probability, $G(k|\theta)$. One might conclude that the number of degrees of freedom scales like $f = N$. In this case, the Pólya MEP is H plus cross-entropy terms. If γ is sufficiently small, this is indeed true and the Pólya processes essentially behave like Bernoulli processes. If γ is sufficiently large, however, the Pólya process is likely to enter the WTA dynamics if one state gets sampled repeatedly in the very beginning of the process. How often on average do we expect state i to be sampled in a

row at the beginning of a Pólya process? The answer is

$$\langle n\rangle(q_i) = \sum_{n=0}^{\infty} n\left(1 - \frac{q_i + \gamma n}{1 + \gamma n}\right) \prod_{m=0}^{n-1} \frac{q_i + \gamma n}{1 + \gamma n}. \quad (26)$$

To first order in γ, one can estimate that

$$\langle n\rangle(q_i) \sim \frac{q_i}{1 - q_i - \kappa(q_i)\gamma}, \quad (27)$$

where $\kappa(q_i) > 1 - q_i$. As $\gamma \to (1 - q_i)/\kappa(q_i)$ from below, $\langle n\rangle(q_i) \to \infty$. This means that if states j violate the condition $\gamma < (1 - q_j)/\kappa(q_j) \leqslant 1$, it becomes likely that the Pólya process enters the WTA dynamics. Practically this means that usually WTA behavior can be observed if a state i gets sampled repeatedly within the first few steps of the Pólya process. Otherwise, the effective reinforcement γ' becomes too small for the process to enter the WTA dynamics, and the sampling distribution $q(N)$ approaches that of a Bernoulli process.

For sufficiently large γ, one finds the situation in which $G(k|\theta)$ can be written as $G(k|\theta) = \tilde{M}(k)\tilde{G}(k|\theta)/M(k)$, so that $MG = \tilde{M}\tilde{G}$ [19]. This means that the probability for the histogram $P = \tilde{M}\tilde{G}$ no longer depends on the multinomial factor M at all. One observes that for $\gamma > 0$, the expression $\log \tilde{M}$ scales very differently than multinomial multiplicities. With $f = 1$ the MEP entropy $S_{\text{MEP}} \equiv \frac{1}{f} \log \tilde{M}$ becomes a well-defined *generalized* relative entropy, and $S_{\text{cross}} = -\frac{1}{f} \log \tilde{G}$ is a *generalized* cross-entropy functional. In [19] we have shown in detail that

$$S_{\text{MEP}}(p|N) \sim -\sum_{i=1}^{W} \log(p_i + 1/N),$$

$$S_{\text{cross}}(p|q,\gamma,N) \sim -\frac{1}{\gamma} \sum_{i=1}^{W} q_i \log(p_i + 1/N). \quad (28)$$

The numerical values for the WTA dynamics (one winner and $W - 1$ losers) are

$$S_{\text{MEP}} \sim (W - 1) \log N + \text{const},$$

$$S_{\text{cross}} \sim \frac{1 - q_j}{\gamma} \log N + \text{const}. \quad (29)$$

The generalized relative entropy S_{rel} can also be viewed as the information divergence of Pólya processes,

$$S_{\text{rel}} = D_{\text{Pólya}}(p|\theta) = \sum_{i=1}^{W} \left(\frac{q_i}{\gamma} - 1\right) \log(p_i + 1/N). \quad (30)$$

$D_{\text{Pólya}}$ is convex in p_i only if $\gamma < q_i$. The processes becomes unstable if the reinforcement parameter γ is sufficiently large. This intrinsic instability of self-reinforcing processes makes MEP predictions of the distribution function $p = (p_1, \ldots, p_W)$ unreliable since large deviations from the maximum configuration p^* remain probable, even for large N. In other words, no well-defined typical sets of paths x form with respect to the distribution of states $i \in \Omega$. However, quite remarkably, ensembles of Pólya urns show stable frequency and rank distributions. If we want to predict the relative frequencies of states ordered according to their rank, the largest frequency having rank $r = 1$, the second largest frequency rank $r = 2$, etc., then this rank distribution $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_W)$ can still be predicted with high accuracy [19]. Pólya urn paths produce typical sets with respect to the most likely observed rank distribution.

E. Summary Pólya urn processes

Pólya urn processes either enter WTA dynamics or behave as a Bernoulli process. For WTA scenarios, one finds that $S_{\text{IT}} \sim \frac{1}{N} \log N$, the extensive entropy is $S_{\text{EXT}} = S_{c,0}$, where $c < 0$, and $S_{\text{MEP}} \sim (W - 1) \log N$ (see Table I). The corresponding numerical values of the different entropies yield similar results, i.e., Pólya urns,

$$N S_{\text{IT}} \sim S_{\text{cross}} \sim \frac{1 - q_j}{(W - 1)\gamma} S_{\text{MEP}}. \quad (31)$$

Again, S_{cross} is a measure of information production. However, instead of measuring the information rate, which becomes zero, it measures the *total* information production. This matches the intuition that the most likely "winner" is a state that happens to be "in the lead" at the very beginning of the process. With some nonvanishing probability, another state can take over the lead within the first few steps. However, if this happens it becomes very unlikely that the Pólya urn process can still enter the WTA dynamics because of the decreasing effective reinforcement parameter γ'. The process then asymptotically approaches a Bernoulli process, where the three entropies are degenerate, $S_{\text{IT}} \sim S_{\text{EXT}} \sim S_{\text{MEP}} \sim H$.

IV. THE THREE ENTROPIES OF SAMPLE-SPACE-REDUCING PROCESSES

A. Sample-space-reducing processes

Sample-space-reducing (SSR) processes are processes whose sample space reduces as they evolve over time. They provide a way to explain the origin and ubiquity of power laws in complex systems, and Zipf's law in particular [22,23]. SSR processes are typically irreversible, dissipative processes that are driven between sources and sinks. Complicated

TABLE I. Extensive entropy, information theoretic entropy rate, and maximum entropy for Pólya, sample-space-reducing, and multinomial mixture processes. $H(p)$ is defined in Eq. (2) and $f(q)$ is the mixing kernel. Expressions are generally valid for large N and W.

	Pólya process (WTA)	SSR process	Multinomial mixture process
S_{EXT}	$S_{1-\frac{\gamma}{1-q_j},0}$	$S_{1,1} = H(p)$ (ensemble)	$S_{1,1} = H(p)$
S_{IT}	$\frac{1-q_j}{\gamma} \frac{1}{N} \log N$	$1 + \frac{1}{2} \log W$	$\int_0^1 dq\, f(q)\, H(q)$
S_{MEP}	$-\sum_i \log p_i$	$-\sum_{i=2}^{W} [p_i \log(\frac{p_i}{p_1}) + (p_1 - p_i) \log(1 - \frac{p_i}{p_1})]$	depends on mixing kernal $f(q) = \mu(q)\gamma(q|\theta) \Rightarrow S_{\text{MEP}} = \log(\mu(q))$

driven dissipative processes such as sandpile dynamics [24], can often be decomposed into simpler SSR processes. Examples of sample-space-reducing processes include fragmentation processes, sentence formation [23], diffusion and search processes on networks [25], and cascading processes [26].

SSR processes can be viewed as processes in which the currently occupied state determines the sample space for the next. If the system is in state i, it can sample states from a sample space Ω_i. Often sample spaces are nested along the process, meaning that $\Omega_i \subset \Omega_j \Leftrightarrow i > j$. In such cases, as the process evolves, the sample space successively becomes smaller. Eventually, a SSR process ends in a sink state, $i = 1$ (Ω_1 is the empty set). The dynamics of such systems is irreversible and nonergodic. To keep the dynamics going, SSR processes have to be restarted, which can lead to a stationary, driven, and irreversible process that is effectively ergodic.

A simple way to depict a SSR process is a ball bouncing downward random distances on a staircase. It never jumps upward. Each stair represents a state i. State $i = 1$ corresponds to the bottom, while state $i = W$ corresponds to the top of the staircase; see Fig. 2(a). Obviously, successive sample spaces are nested. A ball on step i can sample from all steps below itself $j < i$ with equal or prior probabilities q. If the steps carry prior probabilities $q = (q_1, \ldots, q_W)$ [which can be intuitively interpreted as the widths of the steps, Fig. 2(b)], the process will visit state $j < i$ with probability q_j/Q_{i-1}, where $Q_i = \sum_{s=1}^{i} q_s$ is the cumulative distribution of q up to i. Regardless of q (exceptions are discussed in [25]), the SSR processes still follows Zipf's law in the visiting distributions, $p_i = i^{-1}$. By restarting the process, one forces the process to become quasiergodic, meaning that a stationary distribution p exists, despite the process being irreversible. By allowing the process to jump to any position with a given frequency $1 - \lambda$, the visiting distributions remain exact power laws, $p_i = i^{-\lambda}$ [22,25]. In the following, we discuss the three entropies of the "staircase process."

B. The information rate of SSR processes

Note that SSR processes are Markov processes, and the probability of sampling x_n only depends on the previous sample x_{n-1}. Considering ensembles of "staircases" (restarting the SSR process every time it stops) allows us to treat the process as if it were ergodic, and well-defined asymptotic distributions $p = (p_1, \ldots, p_W)$ exist. The entropy production of typical sequences therefore yields

$$S_{\mathrm{IT}}(x) = -\frac{1}{N}\sum_{n=1}^{N} \log p(x_n|x_{n-1}) \sim -\sum_{i,j=1}^{W} p(j|i)p_i \log p(j|i) = \sum_{i=1}^{W} p_i H_i. \quad (32)$$

The entropy production of the SSR process is given by the *conditional entropy*. For uniform priors $q_i = 1/W$ one computes the numerical value of the SSR entropy production,

$$S_{\mathrm{IT}} = p_1 \log W + p_1 \sum_{i=2}^{W} \frac{1}{i} \log i \sim 1 + \frac{1}{2}\log W + O(1/\log(W)). \quad (33)$$

Here we replaced sums $\sum_{i=a}^{b} f(i)$ by integrals $\int_{a-1/2}^{b+1/2} dx\, f(x)$. Note that the 1 in Eq. (33) arises from the restarting procedure.

C. The extensive entropy of SSR processes

We quantify how the phase space of a SSR process grows by the number of decisions (where the ball jumps next) the process takes along its path. A Bernoulli process on W states chooses between W possible successor states at every time

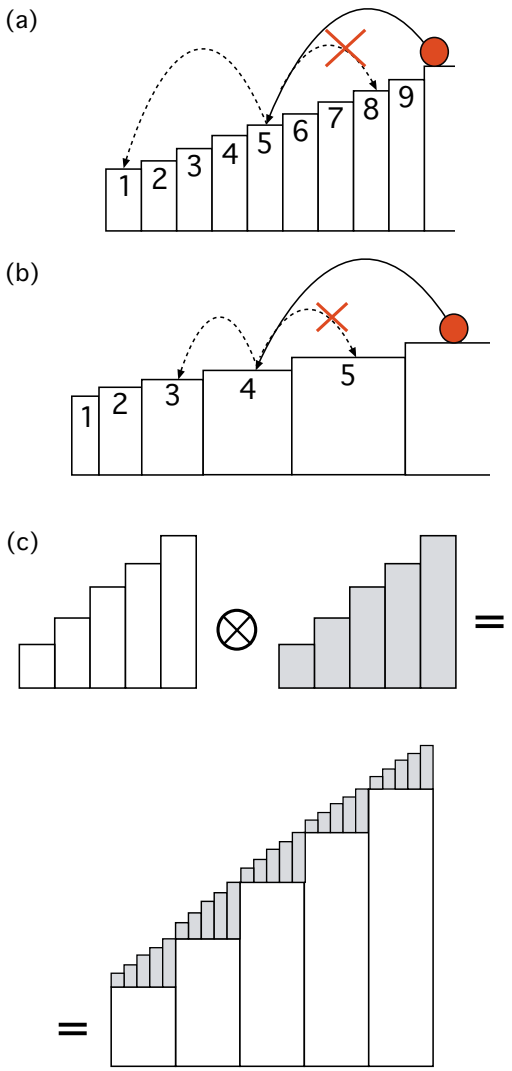

FIG. 2. (a) Pictorial view of a SSR process. A ball bounces downward only, with random step sizes. After several iterations of the process, the visiting probabilities of states i approach $p_i = i^{-1}$ (Zipf's law). (b) SSR with nonuniform prior probabilities. For a wide class of prior probabilities, the visiting distributions still follow Zipf's law. (c) Combining two staircase processes through a "Cartesian product."

step. After N samples, the process selects one specific path among the W^N possible.

The effective number of decisions in a SSR process is computed using Eq. (7). Note that by restarting the SSR process it becomes quasiergodic, and that each state is visited with probability $p_i = p_1/i$, with $1/p_1 = \sum_{i=1}^{W} 1/i$. At state $i > 1$, the process can sample from $W_i = i - 1$ states, and restarting the process once it hits state $i = 1$ means $W_1 = W$ (it can jump anywhere). With this we compute the typical size of the phase space,

$$\hat{W}(N) \equiv \prod_{n=1}^{N} W_{x_n} \sim \prod_{i=1}^{W} W_i^{p_i N} = \bar{W}^N. \tag{34}$$

Consequently, $\bar{W} = W^{p_1} \prod_{i=2}^{W} (i-1)^{p_i}$, and the average number of choices per step involved in sampling a typical SSR sequence x is given by the numerical value

$$\log \bar{W} = \tfrac{1}{2} \log W + 1 + O(1/\log W). \tag{35}$$

The contribution of the constant 1 comes from restarting the process. This implies that $\bar{W} \sim e\sqrt{W} > 1$.

The definition of extensivity is closely related to the way systems are composed. Staircase A with $W(A)$ states can be combined with staircase B with $W(B)$ steps, and to a staircase AB by substituting each step of staircase A with a copy of staircase B; see Fig. 2(c). We get

$$\bar{W}(AB) = e\sqrt{W(A)W(B)} = \frac{1}{e}\bar{W}(A)\bar{W}(B). \tag{36}$$

If we compose staircase A N times with itself, we get $\bar{W}(A(N)) = e(\bar{W}(A)/e)^N$. In other words, the quasiergodic SSR has an exponentially growing phase space, and the extensive entropy is given by $S_{\text{EXT}} = H$.

D. The MEP of SSR processes

To arrive at the MEP for SSR processes X_{SSR} with histogram $k = (k_1, \ldots, k_W)$ as the macrostate, we need to determine the probability $P(k|q) = M(k)G(k|q)$ after N observations of the process and determine the maximum configuration k^* that maximizes $P(k|q)$. To compute M, we first decompose any sampled sequence $x = (x_1, \ldots, x_N)$ into shorter sequences x^r, such that $x = x^1 x^2 \cdots x^R$ is a concatenation of such shorter sequences. Any sequence x^r is a sample of executing X until X stops. We refer to x^r as one "run" of X. This means that any run $x^r = x_1^r x_2^r \cdots x_{N_r}^r$ is a monotonously decreasing sequence of states, $x_n^r > x_{n+1}^r$, ending in $x_{N_r}^r = 1$, where X stops and needs to be restarted. Note that $\sum_{r=1}^{R} N_r = N$. Since every run ends in state 1, the number of runs equals the number of times state 1 is sampled, $R = k_1$. Arranging x in a table with W columns and k_1 rows, and denoting a stair that gets visited by $*$ and a stair that does not get visited within a run by $-$, allows us to determine the probability G and the multiplicity M of a sequence x,

$r \times i$	W	$W-1$	$W-2$	$\cdots$	2	1
1	$*$	$-$	$-$	$\cdots$	$*$	$*$
2	$-$	$*$	$*$	$\cdots$	$-$	$*$
3	$*$	$-$	$*$	$\cdots$	$-$	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$R-2$	$-$	$*$	$*$	$\cdots$	$-$	$*$
$R-1$	$-$	$*$	$-$	$\cdots$	$*$	$*$
R	$-$	$-$	$*$	$\cdots$	$-$	$*$
	k_W	k_{W-1}	k_{W-2}	$\cdots$	k_2	k_1

(37)

We can directly assess the number M of sequences x that have the same histogram k. Note that column i is $k_1 = R$ entries long and contains k_i items; $k_i \leqslant k_1$. Therefore, one can produce all those sequences x by rearranging k_i visits to state i in a column of k_1 possible positions. Each column $i > 1$ therefore contributes to M with the binomial factor $\binom{k_1}{k_i} = k_1!/k_i!/(k_1 - k_i)!$. As a consequence, one finds $M(k) = \prod_{i=2}^{W} \binom{k_1}{k_i}$, and the reduced MEP entropy $\frac{1}{N} \log M$ is given by

$$S_{\text{MEP}} = -\sum_{i=2}^{W} \left[p_i \log\left(\frac{p_i}{p_1}\right) + (p_1 - p_i) \log\left(1 - \frac{p_i}{p_1}\right) \right]. \tag{38}$$

The numerical values are

$$\begin{aligned} S_{\text{MEP}} &= p_1 \sum_{i=2}^{W} \left(1 - \frac{1}{i}\right) \log\left(1 - \frac{1}{i}\right) + p_1 \sum_{i=2}^{W} \frac{1}{i} \log i \\ &\sim \frac{1}{2} \log W + 1 + O(1/\log W). \end{aligned} \tag{39}$$

Similarly, one can determine the probability of sampling a particular sequence x. Each visit to a state $i > 1$ in the sequence x contributes to the probability of the next visit to a state $j < i$ with a factor $1/Q_{i-1}$, whatever j gets sampled. Only if $i = 1$ do we not get such a renormalization factor, since the process restarts and all states i are valid targets with probability q_i. It follows that $G(k|q,N) = \prod_{i=1}^{W} q_i^{k_i} \prod_{j=2}^{W} Q_{i-1}^{-k_i}$, and the cross-entropy is found to be

$$S_{\text{cross}}(p|q) = -\sum_{i=1}^{W} p_i \log q_i + \sum_{i=2}^{W} p_i \log Q_{i-1}. \tag{40}$$

Since terms in S_{cross} do not cancel terms in S_{MEP}, we can safely identify S_{MEP} with the reduced Boltzmann entropy, $S_{\text{MEP}} = s_B$.

The relative entropy of the staircase process is

$$S_{\text{rel}} = S_{\text{cross}} - S_{\text{MEP}}. \tag{41}$$

To get the maximum configuration, we have to minimize S_{rel} with respect to p under the constraint $\sum_{i=1}^{W} p_i = 1$. The result is derived in Appendix B and reads

$$p_i = p_1 \frac{q_i}{Q_i}. \tag{42}$$

For constant prior probabilities $q_i = 1/W$, this yields Zipf's law $p_i = p_1/i$, with p_1 a normalization constant.

Note that the form of the MEP entropy of SSR processes, $S_{\rm MEP}(p) = \sum_{i=2}^{W} s(p_1, p_i)$, is not of trace form, since the state $i = 1$ remains entangled with every other state $j > 1$. SSR processes violate almost all the SK axioms. For perfectly ordered states with distributions $p_i = \delta_{ij}$ that are concentrated on a single state j, $S_{\rm MEP}(\delta_{ij}) = 0$. For the uniform distribution $p_i = 1/W$, we get $S_{\rm MEP}(p) = 0$. This property has been advocated by Gell-Mann and Lloyd for functionals measuring a so-called *effective complexity* [27,28]. This property emerges from the fact that for SSR processes, the uniform distribution can only be obtained if the process evolves along the particular sequence $W \to W-1 \to W-2 \to \cdots \to 2 \to 1$, which is immensely unlikely.

E. Summary SSR processes

For sufficiently large W, the values for entropy production $S_{\rm IT}$, of MEP entropy $S_{\rm MEP}$ (reduced Boltzmann entropy s_B), and the generalized cross-entropy $S_{\rm cross}$, all yield the same numerical values,

$$S_{\rm IT} \sim S_{\rm MEP} \sim S_{\rm cross} \sim \tfrac{1}{2} \log W + O(1 + 1/\log W). \quad (43)$$

Much of what is true for Markov processes remains true for SSR processes, which become Markovian by restarting the process once it stops in $i = 1$. The reduced Boltzmann entropy again measures the typical information rate of the process and determines the amount of information that is required to optimally code typical SSR processes. Comparing Eq. (43) with entropy production of Bernoulli processes $\log W$, note that typical SSR processes only need half the information for encoding a message. It is remarkable that SSR processes, as driven dissipative systems, show enhanced compressibility.

V. MULTINOMIAL MIXTURE PROCESSES

Multinomial mixture processes (MMPs) can be viewed as two-step processes, where an urn is filled with dice with W faces. Each die may have individual biases $q = (q_1, \ldots, q_W)$. From this urn we draw a die, toss it, record the outcome, and put it back into the urn. In other words, one draws dice with biases q according to some fixed probability density function $f(q)$ that is called the *mixing kernel*. Assume f to be sufficiently smooth and nonvanishing for all states i.

A. Entropy production and extensive entropy of multinomial mixture processes

The MMP samples from the states $i = 1, \ldots, W$ again and again. The process is stationary, and if f is smooth, then $\bar{W} > 1$. As a consequence, the extensive entropy of such processes must be $(c,d) = (1,1)$,

$$S_{\rm EXT}(p) = S_{(1,1)}(p) = H(p), \quad (44)$$

meaning that the extensive entropy is H.

MMPs are ergodic. Therefore, for each set of biases q in the mixture, one gets a typical contribution $H(q)$ to the entropy production, and the entropy rate of a typical sequence is given by the expectation value,

$$S_{\rm IT}(x) \sim \int_0^1 dq\, f(q)\delta(1 - |q|_1)H(q) \equiv \langle H \rangle_f, \quad (45)$$

which is merely the conditional entropy to draw a die with weights q, given that q is drawn with probability f. Note that the expected frequencies are

$$p_i = \int_0^1 dq\, f(q)\delta(1 - |q|_1)q_i \equiv \langle q_i \rangle_f. \quad (46)$$

It follows that in general $H(\langle q \rangle_f) > \langle H \rangle_f$, meaning that Shannon entropy of the stationary distribution overestimates the information rate of the process.

B. The MEP of multinomial mixture processes

Assume a MMP with $\theta = q$. The probability to sample histogram k is

$$P(k) = M(k) \int_0^1 dq\, f(q)\delta(|q|_1 - 1) \prod_{j=1}^{W} q_j^{k_j}, \quad (47)$$

where $M(k)$ is the multinomial factor, $|q|_1 = \sum_{i=1}^{W} q_i$, and f is normalized, $1 = \int_0^1 dq\, f(q)\delta(|q|_1 - 1)$. Just as in the case of the Pólya process, one might naively think that the MEP functional is H plus cross-entropy terms. Again, this turns out to be wrong. Consider the identity

$$\left[M(k) \frac{(N+W-1)!}{N!} \right]^{-1} = \int_0^1 dq \prod_{i=1}^{W} q_i^{k_i} \delta(|q|_1 - 1). \quad (48)$$

Since for a distribution p with $|p|_1 = 1$ the function $\prod_{i=1}^{W} q_i^{p_i}$ is maximal for $p = q$, we see that for large N,

$$\prod_{i=1}^{W} \delta\left(q_i - \frac{k_i}{N}\right) \sim M(k) \frac{(N+W-1)!}{N!} \prod_{i=1}^{W} q_i^{k_i} \delta(|y|_1 - 1) \quad (49)$$

forms a so-called δ sequence. Inserting Eq. (49) into Eq. (47) gives

$$P(k) \sim N^{1-W} f\left(\frac{k}{N}\right), \quad (50)$$

and the relative entropy of MMPs with $f(q|\theta)$ is

$$S_{\rm rel}(p|\theta) = -\log f(p|\theta). \quad (51)$$

If $S_{\rm rel}$ can be decomposed into $S_{\rm rel} = S_{\rm MEP} - \tilde{S}_{\rm cross}$, it depends on the mixing kernel f. If it factorizes $f(q|\theta) = \mu(q)\gamma(q|\theta)$, then $S_{\rm MEP}(q) \sim \log \mu(q)$, $S_{\rm cross}(q|\theta) \sim -\log \gamma(q|\theta)$, and

$$P(k|\theta) = M(k) \int dq\, \delta(|q|_1 - 1) \prod_{j=1}^{W} q_j^{k_j} \frac{1}{Z} e^{S_{\rm MEP} - S_{\rm cross}}, \quad (52)$$

where Z is a normalization constant.

VI. CONCLUSIONS

For simple systems, the concepts of thermodynamic entropy, information-theoretic entropy, and the entropy in the maximum entropy principle all lead to the same entropy functional $H(p)$; it is degenerate. The essence behind simple systems and processes rests in the fact that they are all basically built on multinomial Bernoulli processes. We showed that Bernoulli processes generically lead to H regardless

of the entropy concept that is used. We showed in three concrete examples that this degeneracy is broken for more complex processes, and that the three entropy concepts lead to completely distinct functional forms. The entropy concepts now capture information about distinct properties of the underlying system. The three processes studied were the Pólya process as an example of a self-reinforcing process, sample-space-reducing processes as an example of history-dependent processes with power-law distribution functions, and multinomial mixture processes, which serve as an example of composed stochastic processes. The results are summarized in Table I. The processes discussed here are relatively simple when compared to stochastic processes that occur in actual nonergodic complex adaptive systems, which often are self-reinforcing, path-dependent, and composed of multiple dynamics. The main contribution of our exercise here is that it shows unambiguously that for any process that cannot be based on, or be traced back to, Bernoulli processes, one needs to exactly specify which concept of entropy one is talking about before it makes sense to try to compute it. In general, the three concepts have to be computed system class by system class. The naive use of the expression H as a one-size-fits-all concept will inevitably lead to confusion and nonsense. It remains to be seen if systems and processes can be classified into families that share the same three faces of entropy.

ACKNOWLEDGMENTS

This work was supported by the Austrian Science Foundation FWF under Projects No. P29032 and No. I3073.

APPENDIX A: EXISTENCE OF UNIQUE EXTENSIVE ENTROPY FOR NONEXTENSIVE SYSTEMS

Assume that the effective phase-space volume is given by

$$\hat{W}(N) \equiv \prod_{n=1}^{N} \bar{W}(X_n). \tag{A1}$$

Since $\hat{W}(N)$ is monotonically increasing in N, an inverse function L_X exists such that $L_X(\hat{W}(N)) = N$, and a unique extensive trace-form functional can be found,

$$S_{\mathrm{EXT}}(p) = \sum_{x\in\Omega^N} s(p(x)). \tag{A2}$$

Here $q(x)$ is the probability to sample path x, such that for sequences $x(N)$ one obtains

$$S_{\mathrm{EXT}}(q(x(N))) = N s_0, \tag{A3}$$

with $s_0 = \tilde{W}(1)s(1/\tilde{W}(1))$. If we look at a reference process, where the path probabilities $q(x)$ are uniformly concentrated on $\hat{W}(N)$ paths, it follows that

$$\sum_{x\in\Omega^N} s(q(x)) \sim \hat{W}(N)s\left(\frac{1}{\hat{W}(N)}\right). \tag{A4}$$

Clearly, $\hat{W}(N)s(\frac{1}{\hat{W}(N)}) = N s_0$ is exactly solved by

$$s(x) = s_0\, x\, L_X\left(\frac{1}{x}\right). \tag{A5}$$

APPENDIX B: SOLVING THE MEP FOR SSR PROCESSES

To maximize the MEP of the staircase process, Eq. (41), with respect to the probabilities p under the constraint $\sum_{i=1}^{W} p_i = 1$, where W is the number of possible states, we may proceed as follows. The staircase MEP requires us to solve $\delta(\psi(p|q,N) - \alpha(\sum_{i=1}^{W} p_i - 1)) = 0$, where $\psi = S_{\mathrm{MEP}} - S_{\mathrm{cross}} = -S_{\mathrm{rel}}$ of the SSR process and α is the Lagrange multiplier guaranteeing the constraint. This means that every derivative of the constrained functional with respect to p_i must be zero. For $i > 1$ one gets

$$p_i = \frac{p_1}{1 + \zeta \frac{Q_{i-1}}{q_i}}, \tag{B1}$$

where $\zeta = \exp(\alpha)$. Similarly, for $i = 1$ one finds

$$q_1 = \zeta \exp\left[\sum_{i=2}^{W} \log\left(1 - \frac{p_i}{p_1}\right)\right]. \tag{B2}$$

Solving these two equations self-consistently, one finds (at least numerically) that $\zeta = 1$, and the solution of the MEP is

$$p_i = p_1 \frac{q_i}{Q_i}. \tag{B3}$$

[1] S. Carnot, *Reflexions sur la Puissance Motrice du Feu et sur les Machines Propres a Developper cette Puissance* (Bachelier, Paris, 1824).
[2] R. Clausius, Ann. Phys. Chem. **169**, 481 (1854).
[3] C. Kittel, *Elementary Statistical Physics* (Wiley, New York, 1958).
[4] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[5] L. G. Kraft, M.Sc. thesis, MIT, 1949.
[6] B. McMillan, IEEE Trans. Inf. Theor. **2**, 115 (1956).
[7] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[8] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[9] B. McMillan, Ann. Math. Stat. **24**, 196 (1953).
[10] L. Breiman, Ann. Math. Stat. **28**, 809 (1957); **31**, 809(E) (1960).
[11] N. G. van Kampen, in *Essays in Theoretical Physics in Honour of Dirk ter Haar*, edited by W. E. Parry (Pergamon, Oxford, 1984), pp. 303–312.
[12] R. Hanel and S. Thurner, Europhys. Lett. **96**, 50003 (2011).
[13] R. Hanel, S. Thurner, and M. Gell-Mann, Proc. Natl. Acad. Sci. (USA) **111**, 6905 (2014).
[14] R. Hanel and S. Thurner, Entropy **15**, 5324 (2013).
[15] R. Hanel and S. Thurner, Europhys. Lett. **93**, 20006 (2011).
[16] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics* (Springer, New York, 2009).

[17] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[18] M. Planck, Ann. Phys. **309**, 553 (1901).
[19] R. Hanel, B. Corominas-Murtra, and S. Thurner, New J. Phys. **19**, 033008 (2017).
[20] F. Eggenberger and G. Pólya, Z. Angew. Math. Mech. **3**, 279 (1923).
[21] G. Pólya, Ann. Inst. Henri Poincare **1**, 117 (1930).
[22] B. Corominas-Murtra, R. Hanel, and S. Thurner, Proc. Natl. Acad. Sci. (USA) **112**, 5348 (2015).
[23] S. Thurner, R. Hanel, and B. Corominas-Murtra, J. R. Soc. Interface **12**, 20150330 (2015).
[24] A. Corral, Phys. Rev. E **69**, 026107 (2004).
[25] B. Corominas-Murtra, R. Hanel, and S. Thurner, New J. Phys. **18**, 093010 (2016).
[26] B. Corominas-Murtra, R. Hanel, and S. Thurner, Sci. Rep. **7**, 11223 (2017).
[27] M. Gell-Mann and S. Lloyd, *Effective Complexity*, SFI working paper 387 (2003-12-068).
[28] M. Gell-Mann and S. Lloyd, Complexity **2**, 44 (1996).