How multiplicity determines entropy and the derivation of the maximum entropy principle for complex systems

Rudolf Hanel[a], Stefan Thurner[a,b,c], and Murray Gell-Mann[b,1]

[a]Section for Science of Complex Systems, Medical University of Vienna, 1090 Vienna, Austria; [b]Santa Fe Institute, Santa Fe, NM 87501; and [c]International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria

Contributed by Murray Gell-Mann, April 4, 2014 (sent for review January 30, 2014)

The maximum entropy principle (MEP) is a method for obtaining the most likely distribution functions of observables from statistical systems by maximizing entropy under constraints. The MEP has found hundreds of applications in ergodic and Markovian systems in statistical mechanics, information theory, and statistics. For several decades there has been an ongoing controversy over whether the notion of the maximum entropy principle can be extended in a meaningful way to nonextensive, nonergodic, and complex statistical systems and processes. In this paper we start by reviewing how Boltzmann–Gibbs–Shannon entropy is related to multiplicities of independent random processes. We then show how the relaxation of independence naturally leads to the most general entropies that are compatible with the first three Shannon–Khinchin axioms, the (*c*,*d*)-entropies. We demonstrate that the MEP is a perfectly consistent concept for nonergodic and complex statistical systems if their relative entropy can be factored into a generalized multiplicity and a constraint term. The problem of finding such a factorization reduces to finding an appropriate representation of relative entropy in a linear basis. In a particular example we show that path-dependent random processes with memory naturally require specific generalized entropies. The example is to our knowledge the first exact derivation of a generalized entropy from the microscopic properties of a path-dependent random process.

thermodynamics | out-of-equilibrium process | driven systems | random walk

Many statistical systems can be characterized by a macrostate for which many microconfigurations exist that are compatible with it. The number of configurations associated with the macrostate is called the phase-space volume or multiplicity, M. Boltzmann entropy is the logarithm of the multiplicity,

$$S_B = k_B \log M, \qquad [1]$$

and has the same properties as the thermodynamic (Clausius) entropy for systems such as the ideal gas (1). We set $k_B = 1$. Boltzmann entropy scales with the degrees of freedom f of the system. For example, for N noninteracting point particles in three dimensions, $f(N) = 3N$. Systems where S_B scales with system size are called extensive. The entropy per degree of freedom $s_B = (1/f)S_B$ is a system-specific constant. Many complex systems are nonextensive, meaning that if two initially insulated systems A and B, with multiplicities M_A and M_B, respectively, are brought into contact, the multiplicity of the combined system is $M_{A+B} < M_A M_B$. For such systems, which are typically strongly interacting, non-Markovian, or nonergodic, S_B and the effective degrees of freedom $f(N)$ do no longer scale as N. Given the appropriate scaling for $f(N)$, the entropy s_B is a finite and nonzero constant in the thermodynamic limit, $N \to \infty$.

A crucial observation in statistical mechanics is that the distribution of all macrostate variables gets sharply peaked and narrow as system size N increases. The reason behind this is that the multiplicities for particular macrostates grow much faster with N than those for other states. In the limit $N \to \infty$ the probability of measuring a macrostate becomes a Dirac delta, which implies that one can replace the expectation value of a macrovariable by its most likely value. This is equivalent to maximizing the entropy in Eq. **1** with respect to the macrostate. By maximizing entropy one identifies the "typical" microconfigurations compatible with the macrostate. This typical region of phase space dominates all other possibilities and therefore characterizes the system. Probability distributions associated with these typical microconfigurations can be obtained in a constructive way by the maximum entropy principle (MEP), which is closely related to the question of finding the most likely distribution functions (histograms) for a given system.

We demonstrate the MEP in the example of coin tossing. Consider a sequence of N independent outcomes of coin tosses, $x = (x_1, x_2, \ldots, x_N)$, where x_i is either head or tail. The sequence x contains k_1 heads and k_2 tails. The probability of finding a sequence with exactly k_1 heads and k_2 tails is

$$P(k_1, k_2|\theta_1, \theta_2) = \binom{N}{k_1} \theta_1^{k_1} \theta_2^{k_2} = M^{\text{bin}}(k) G(k|\theta), \qquad [2]$$

where $M^{\text{bin}}(k) \equiv \binom{N}{k_1}$ is the binomial factor. We use the shorthand notation $k = (k_1, k_2)$ for the histogram of k_1 heads and k_2 tails and $\theta = (\theta_1, \theta_2)$ for the marginal probabilities for throwing head or tail. For the relative frequencies $p_i \equiv k_i/N$ we write $p = (p_1, p_2)$. We also refer to θ as the "biases" of the system. The probability of observing a particular sequence x with histogram k is given by $G(k|\theta) \equiv \theta_1^{k_1} \theta_2^{k_2}$. It is invariant under permutations of the sequence x because the coin tosses are independent. All possible sequences x with the same histogram k have identical

Significance

The maximum entropy principle (MEP) states that for many statistical systems the entropy that is associated with an observed distribution function is a maximum, given that prior information is taken into account appropriately. Usually systems where the MEP applies are simple systems, such as gases and independent processes. The MEP has found thousands of practical applications. Whether a MEP holds for complex systems, where elements interact strongly and have memory and path dependence, remained unclear over the past half century. Here we prove that a MEP indeed exists for complex systems and derive the generalized entropy. We find that it belongs to the class of the recently proposed (*c*,*d*)-entropies. The practical use of the formalism is shown for a path-dependent random walk.

Author contributions: R.H., S.T., and M.G.-M. designed research, performed research, contributed new reagents/analytic tools, and wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: mgm@santafe.edu.

probabilities. $M^{\text{bin}}(k)$ is the respective multiplicity, representing the number of possibilities to throw exactly k_1 heads and k_2 tails. As a consequence Eq. **2** becomes the probability of finding the distribution function p of relative frequencies for a given N. The MEP is used to find the most likely p. We denote the most likely histogram by $k^*(\theta,N)$ and the most likely relative frequencies by $p^*(\theta,N)=k^*(\theta,N)/N$.

We now identify the two components that are necessary for the MEP to hold. The first is that $P(k_1,k_2|\theta_1,\theta_2)$ in Eq. **2** factorizes into a multiplicity $M(k)$ that depends on k only and a factor $G(k|\theta)$ that depends on k and the biases θ. The second necessary component is that the multiplicity is related to an entropy expression. By using Stirling's formula, the multiplicity of Eq. **2** can be trivially rewritten for large N,

$$M^{\text{bin}}(k)=\binom{N}{k_1}\sim e^{N[-p_1\log(p_2)-p_2\log(p_2)]}=e^{NS[p]}, \qquad [3]$$

where an entropy functional of Shannon type (2) appears,

$$S[p]=-\sum_{i=1}^{W=2}p_i\log p_i. \qquad [4]$$

The same arguments hold for multinomial processes with sequences x of N independent trials, where each trial x_n takes one of W possible outcomes (3). In that case the probability for finding a given histogram k is

$$\begin{aligned}&P(k|\theta)=M^{\text{mn}}(k)\theta_1^{k_1}\theta_2^{k_2}\ldots\theta_W^{k_W}=M^{\text{mn}}(k)G(k|\theta),\\ &\text{with}\quad M^{\text{mn}}(k)=\frac{N!}{k_1!k_2!\ldots k_W!}\sim e^{NS[p]}.\end{aligned} \qquad [5]$$

$M^{\text{mn}}(k)$ is the multinomial factor and $S[p]=-\sum_{i=1}^{W}p_i\log(p_i)$. Asymptotically $S[p]=\lim_{N\to\infty}(1/N)\log M^{\text{mn}}(k)$ holds. Extremizing Eq. **5** for fixed N with respect to k yields the most likely histogram, k^*. Taking logarithms on both sides of Eq. **5** gives

$$\underbrace{\frac{1}{N}\log P(k|\theta)}_{-\text{relative entropy}}=\underbrace{\frac{1}{N}\log M^{\text{mn}}(k)}_{S[p]}+\underbrace{\frac{1}{N}\log G(k|\theta)}_{-\text{cross entropy}}. \qquad [6]$$

Obviously, extremizing Eq. **6** leads to the same histogram k^*. The term $-(1/N)\log P(k|\theta)$ in Eq. **6** is sometimes called relative entropy or Kullback–Leibler divergence (4). We identify the first term on the right-hand side of Eq. **6** with Shannon entropy $S[p]$, and the second term is the so-called cross-entropy $-(1/N)\log G(k=pN|\theta)=-\sum_i p_i\log\theta_i$. Eq. **6** states that the cross-entropy is equal to entropy plus the relative entropy. The constraints of the MEP are related to the cross-entropy. For example, let the marginal probabilities θ_i be given by the so-called Boltzmann factor, $\theta_i=\exp(-\alpha-\beta\epsilon_i)$, for the "energy levels" ϵ_i, where β is the inverse temperature and α the normalization constant. Inserting the Boltzmann factor into the cross-entropy, Eq. **6** becomes

$$\frac{1}{N}\log P(k|\theta)=S[p]-\alpha\sum_i p_i-\beta\sum_i p_i\epsilon_i, \qquad [7]$$

which is the MEP in its usual form, where Shannon entropy gets maximized under linear constraints. α and β are the Lagrangian multipliers for the normalization and the "energy" constraint $\sum_i p_i\epsilon_i=U$, respectively. Note that in Eq. **6** we used $f(N)=N$ to scale $\log M^{\text{mn}}(k)$. Any other nonlinear $f(N)$ would yield nonsensical results in the limit of $S[p]$, either 0 or ∞. Comparing $S[p]=\lim_{N\to\infty}(1/N)\log M^{\text{mn}}(k)$ with Eq. **1** shows that indeed, up to a constant multiplicative factor, $s_B=S[p]$. This means that the Boltzmann entropy per degree of freedom of a (uncorrelated) multinomial process is given by a Shannon-type entropy functional.

Many systems that are nonergodic, are strongly correlated, or have long memory will not be of multinomial type, implying that $\hat{P}(x|\theta)$ is not invariant under permutations of a sequence x. For this situation it is not a priori evident that a factorization of $P(k|\theta)$ into a θ-independent multiplicity and a θ-dependent term, as in Eq. **5**, is possible. Under which conditions such a factorization is both feasible and meaningful is discussed in the next section.

Results

When Does a MEP Exist? The Shannon–Khinchin (SK) axioms (2, 5) state requirements that must be fulfilled by any entropy. [Shannon–Khinchin axioms: SK1, entropy is a continuous function of the probabilities p_i only and should not explicitly depend on any other parameters; SK2, entropy is maximal for the equidistribution $p_i=1/W$; SK3, adding a state $W+1$ to a system with $p_{W+1}=0$ does not change the entropy of the system; and SK4, entropy of a system composed of two subsystems, A and B, is $S(A+B)=S(A)+S(B|A)$.] For ergodic systems all four axioms hold. For nonergodic ones the composition axiom (SK4) is explicitly violated, and only the first three (SK1–SK3) hold. If all four axioms hold, the entropy is uniquely determined to be Shannon's; if only the first three axioms hold, the entropy is given by the (c,d)-entropy (6, 7). The SK axioms were formulated in the context of information theory but are also sensible for many physical and complex systems.

The first Shannon–Khinchin axiom (SK1) states that entropy depends on the probabilities p_i only. Multiplicity depends on the histogram $k=pN$ only and must not depend on other parameters. Up to an N-dependent scaling factor the entropy is the logarithm of multiplicity. The scaling factor $f(N)$ removes this remaining N dependence from entropy, so that SK1 is asymptotically fulfilled. In fact, SK1 ensures that the factorization $P(k|\theta)=M(k)G(k|\theta)$ into a θ-independent characteristic multiplicity $M(k)$ and a θ-dependent characteristic probability $G(k|\theta)$ is not arbitrary.

For systems that are not of multinomial nature, we proceed as before: To obtain the most likely distribution function we try to find $k=k^*(\theta,N)$ that maximizes $P(k|\theta)$ for a given N. We denote the generalized relative entropy by

$$D(p|\theta)=-\frac{1}{f(N)}\log P(k|\theta). \qquad [8]$$

Note that whenever an equation relates terms containing k with terms containing p, we always assume $p=k/N$. The maximal distribution $p^*\equiv k^*/N$ therefore minimizes $D(p|\theta)$ and is obtained by solving

$$0=\frac{\partial}{\partial p_i}\left(D(p|\theta)-\alpha\left(\sum_{j=1}^{W}p_i-1\right)\right) \qquad [9]$$

for all $i=1,2,\ldots,W$. α is the Lagrange multiplier for normalization of p.

The histogram $k=(k_1,k_2,\ldots,k_W)$ can be seen as a vector in a W-dimensional space. Let e_i be a W-dimensional vector whose ith component is 1, and all of the others are 0. With this notation the derivative in Eq. **9** can be expressed asymptotically as

$$\frac{\partial}{\partial p_i}D(p|\theta)\sim\frac{N}{f(N)}\log\frac{P(k-e_i|\theta)}{P(k|\theta)}\equiv\frac{N}{f(N)}v_i(k|\theta), \qquad [10]$$

where we write $v_i(k|\theta)$ for the log term. We interpret $v_i(k|\theta)$ as the ith component of a vector $v(k|\theta)\in\mathbb{R}^W$. Let $b_{ji}(k)$ be the ith component of the jth basis vector for any given k; then $v_i(k|\theta)$ has uniquely determined coordinates $c_j(k|\theta)$,

$$v_i(k|\theta)=\sum_{j=1}^{W}c_j(k|\theta)b_{ji}(k). \qquad [11]$$

$v_i(k|\theta)$ has coordinates $c_j(k|\theta)$ in any basis $b_{ji}(k)$. However, as can be easily verified, not all bases are compatible with SK1–SK3 (see condition *i* in the *Theorem* below). The problem of factorizing $P(k|\theta)$ therefore reduces to the problem of finding an appropriate basis. For reasons that become clear below, we choose the following Ansatz for the basis

$$b_{ji}(k) = \frac{\kappa_{ji}}{\gamma_T(N,k_i)} \log \frac{M_{u,T}(k-e_i)u(N)}{M_{u,T}(k)}, \qquad [12]$$

where the functions $M_{u,T}(k)$ are so-called deformed multinomial factors, and κ_{ji} are some appropriately chosen constants. $\gamma_T(N,r) = N[T(r/N) - T((r-1)/N)]$ is a factor depending on a continuous, monotonic, and increasing function T, with $T(0)=0$, and $T(1)=1$. $u(n)$ $(n=0,1,2,\ldots)$ are positive, monotonic increasing functions on the natural numbers. The freedom of choosing κ_{ji}, u, and T in this basis provides a well-defined framework that allows us to derive the conditions for the existence of a MEP. Deformed multinomials are based on deformed factorials that are well known in the mathematical literature (8–13) and are defined as

$$N!_u \equiv \prod_{n=1}^{N} u(n). \qquad [13]$$

For a specific choice of u, deformed multinomials are then defined in a general form as

$$M_{u,T}(k) = \frac{N!_u}{\prod_i \lfloor NT(k_i/N) \rfloor !_u}, \qquad [14]$$

where $\lfloor x \rfloor$ is the largest integer less than x. With the basis of Eq. **12** we can write

$$\begin{aligned} \frac{P(k-e_i|\theta)}{P(k|\theta)} &= \prod_{j=1}^{W} \left(\frac{M_{u,T}(k-e_i)u(N)}{M_{u,T}(k)} \right)^{\left(c_j(k|\theta)/\gamma_T(N,k_i)\right)\kappa_{ji}} \\ &= \prod_{j=1}^{W} u\left(NT\left(\frac{k_i}{N}\right)\right)^{c_j(k|\theta)\kappa_{ji}}. \end{aligned} \qquad [15]$$

Note that this can be done for any process that produces sequences $x=(x_1,x_2,\ldots,x_N)$, where x_n takes one of W values. We can now formulate the following:

Theorem. *Consider the class of processes $x=\{x_n\}_{n=1}^{N}$, with $x_n \in \{1,\ldots,W\}$, parameterized by the biases θ and the number of elements N. The process produces histograms k with probability $P(k|\theta)$. Let N be large and $k^*(\theta,N)$ be the histogram that maximizes $P(k|\theta)$. Assume that a basis of the form given in Eq.* **12** *can be found, for which (i) $\kappa_{1i}=1$, for all $i=1,\ldots,W$, and (ii) for fixed values of N and θ, the coordinate $c_1(k|\theta)$ of $v(k|\theta)$ in this basis, as defined in Eq.* **11**, *becomes a nonzero constant at $k^*(\theta,N)$. [Condition ii means that the first derivatives of $c_1(k|\theta)$ vanish at $k=k^*$ under the condition $\sum k_i = N$, N being constant.] Under these conditions $P(k|\theta)$ factorizes, $P(k|\theta) = M_{u,T}(k)G_{u,T}(k|\theta)$, with*

$$\frac{G_{u,T}(k-e_i|\theta)}{G_{u,T}(k|\theta)} = \prod_{j=2}^{W} u\left(NT\left(\frac{k_i}{N}\right)\right)^{c_j(k|\theta)\kappa_{ji}}. \qquad [16]$$

Moreover, there exists a MEP with generalized entropy $S[p] = (1/f(N))\log M_{u,T}(k)$, for some scaling function $f(N)$. The factors $u(.)^{c_j(k|\theta)\kappa_{ji}}$ in Eq. **16** *represent the constraint terms in the MEP. The solution of the MEP is given by $p^* = k^*/N$.*

The physical meaning of the *Theorem* is that the existence of a MEP can be seen as a geometric property of a given process. This reduces the problem to one of finding an appropriate basis that does not violate axioms SK1–SK3 and that is also convenient. The former is guaranteed by the *Theorem*, and the latter is achieved by using the particular choice of the basis in Eq. **12**.

Condition *ii* of the *Theorem* guarantees the existence of primitive integrals $M_{u,T}(k)$ and $G_{u,T}(k|\theta)$. If condition *i* is violated, the first basis vector b_{1i} of Eq. **12** introduces a functional in p that will in general violate the second Shannon–Khinchin axiom SK2. Conditions *i* and *ii* together determine $S[p]$ up to a multiplicative constant c_1, which can be absorbed in a normalization constant. $G_{u,T}$ may be difficult to construct in practice. However, for solving the MEP it is not necessary to know $G_{u,T}$ explicitly; it is sufficient to know the derivatives of the logarithm for the maximization. These derivatives are obtained simply by taking the logarithm of Eq. **16**. For systems that are compatible with the conditions of the *Theorem*, in analogy to Eq. **6**, a corresponding MEP for the general case of nonmultinomial processes reads

$$\underbrace{\frac{1}{f(N)}\log P(k|\theta)}_{-\text{generalized rel.ent.}} = \underbrace{\frac{1}{f(N)}\log M_{u,T}(k)}_{\text{generalized ent.}S[p]} + \underbrace{\frac{1}{f(N)}\log G_{u,T}(k|\theta)}_{-\text{generalized cross ent.}}. \qquad [17]$$

$f(N)$ has to be chosen such that for large N the generalized relative entropy $D(p|\theta) = -(1/f(N))\log P(k|\theta)$ neither becomes 0 nor diverges for large N. $S[p] = (1/f(N))\log M_{u,T}(k)$ is the generalized entropy, and $C(p|\theta) = -(1/f(N))\log G_{u,T}(k|\theta)$ is the generalized cross-entropy. In complete analogy to the multinomial case, the generalized cross-entropy equals generalized entropy plus generalized relative entropy. Note that in general the generalized cross-entropy $C(p|\theta)$ will not be linear in p_i. In ref. 14 it was shown that the first three Shannon–Khinchin axioms allow only two options for the constraint terms. They can be either linear or of the so-called "escort" type (15), where constraints are given by specific nonlinear functions in p_i (14). No other options are allowed. For the escort case we have shown in refs. 14 and 16 that a duality exists such that the generalized entropy S, in combination with the escort constraint, can be transformed into the dual generalized entropy S^* with a linear constraint. In other words, the nonlinearity in the constraint can literally be subtracted from the cross-entropy and added to the entropy. Compare with the notion of the "corrector" discussed in ref. 17.

The Generalized Entropy. We can now compute the generalized entropy from Eq. **17**,

$$\begin{aligned} S[p] &= \lim_{N\to\infty} f(N)^{-1} \log M_{u,T}(k) \\ &= f(N)^{-1} \left[\sum_{r=1}^{N} \log u(r) - \sum_{i=1}^{W} \sum_{r=1}^{NT(k_i/N)} \log u(r) \right] \\ &= \sum_{r=1}^{N} \frac{1}{N} \frac{N \log u(r)}{f(N)} - \sum_{i=1}^{W} \sum_{r=1}^{NT(p_i)} \frac{1}{N} \frac{N \log u(r)}{f(N)} \\ &= \int_0^1 dy \frac{N \log u(Ny)}{f(N)} - \sum_{i=1}^{W} \int_0^{T(p_i)} dy \frac{N \log u(Ny)}{f(N)} \\ &= -\sum_{i=1}^{W} \int_0^{p_i} dz T'(z) \frac{N \log u(NT(z))}{f(N)} \\ &\quad + \int_0^1 dz T'(z) \frac{N \log u(NT(z))}{f(N)}, \end{aligned} \qquad [18]$$

where $T'(z)$ is the derivative with respect to z. Further, we replace the sum over r by an integral that is correct for large N. The

resulting generalized entropy is clearly of trace form. In refs. 14, 18, and 19 it was shown that the most general form of trace form entropy that is compatible with the first three Shannon–Khinchin axioms is

$$S[p] = -a\left[\sum_{i=1}^{W}\int_0^{p_i} dz\Lambda(z) - \int_0^1 dz\Lambda(z)\right], \qquad [19]$$

where Λ is a so-called generalized logarithm, which is an increasing function with $\Lambda(1)=0$, $\Lambda'(1)=1$; compare refs. 14 and 16. Comparison of the last line of Eq. **18** with Eq. **19** yields the generalized logarithm

$$a\Lambda(z) = T'(z)\frac{N}{f(N)}\log u(NT(z)) - b, \qquad [20]$$

with $a>0$ and b constants. By taking derivatives of Eq. **20**, first with respect to z and then with respect to N, one solves the equation by separation of variables with a separation constant ν. Setting $b=\log\lambda$, we get

$$\begin{aligned}\Lambda(z) &= \frac{T'(z)T(z)^{\nu} - T'(1)}{T''(1)+\nu T'(1)^2}\\ u(N) &= \lambda^{(N^{\nu})}\\ f(N) &= N^{1+\nu}\\ a &= \left(\frac{T''(1)}{T'(1)}+\nu T'(1)\right)\log\lambda.\end{aligned} \qquad [21]$$

By choosing T and ν appropriately one can find examples for all entropies that are allowed by the first three SK axioms, which are the (c,d)-entropies (6, 7). (c,d)-entropies include most trace form entropies that were suggested in the past decades as special cases. The expressions $f(N)$ and $u(x)$ from Eq. **21** can be used in Eqs. **9** and **15** to finally obtain the most likely distribution from the minimum relative entropy,

$$p_i^* = T^{-1}\left(\left[\frac{\log\lambda}{\alpha}\sum_{j=1}^{W} c_j(Np^*|\theta)\kappa_{ji}\right]^{-(1/\nu)}\right), \qquad [22]$$

which must be solved self-consistently. T^{-1} is the inverse function of T. In the case that only the first two basis vectors are relevant (the generalized entropy and one single constraint term), we get distributions of the form

$$p_i^* = T^{-1}\left(\left[1+\nu\left(\hat{\alpha}+\hat{\beta}\epsilon_i\right)\right]^{-\frac{1}{\nu}}\right), \qquad [23]$$

with $\hat{\alpha}=\frac{1}{\nu}\left(\frac{\log\lambda}{\alpha}c_1-1\right)$, $\hat{\beta}=\frac{\log\lambda}{\alpha\nu}c_2(Np^*|\theta)$. In a polynomial basis, specified by $\kappa_{ji}\equiv(i-1)^{j-1}$, the equally spaced "energy levels" are given by $\epsilon_i=(i-1)$. Note that $c_1=1$, and $c_2(p^*N|\theta)$ depends on bias terms.

For a specific example let us specify $T(z)=z$ and $\lambda>1$. Eqs. **21** and **19** yield

$$S[p] = \left(\frac{a}{Q}\right)\frac{1-\sum_{i=1}^{W}p_i^Q}{Q-1},\ [Q\equiv 1+\nu], \qquad [24]$$

which is the so-called Tsallis entropy (20). $\gamma_T(N,r)=1$ for this choice of T. Any other choice of T leads to (c,d)-entropies.

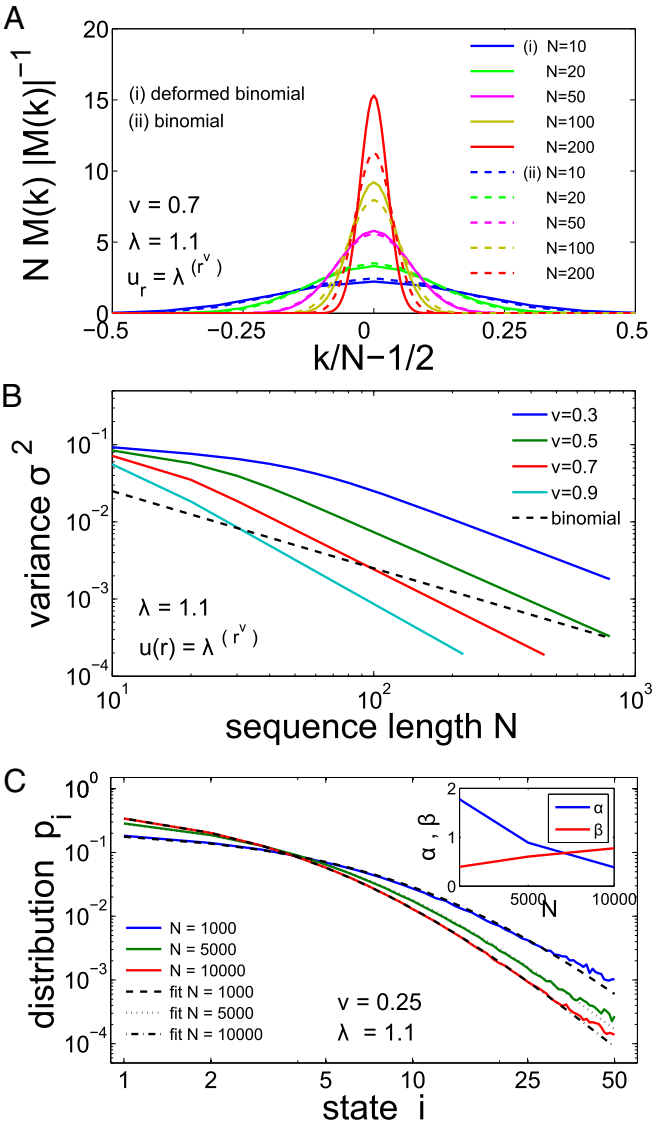

Fig. 1. Numerical results for the path-dependent random process determined by the deformed factorial $N!_u$ with $u_r=(\lambda^{(r^*)}-1)/(\lambda-1)$. (*A*) Normalized generalized binomial factors $M_{u,T}(k_1,N-k_1)$ (solid lines). Distributions get narrower as N increases, which is necessary for the MEP to hold. Dashed lines show the usual binomial factor ($\nu=1$ and $\lambda\to 1$). (*B*) Variance $\sigma^2=\sum_{k_1=0}^{N}M_{u,T}(k_1/N-1/2)^2$ of the normalized generalized binomial factors (solid lines), as a function of sequence length N, for various values ν and $\lambda=1.1$. The dashed line is the variance of the usual binomial multiplicity. (*C*) Probability distributions for the $W=50$ states i from numerical realizations of processes following Eq. **27**, with $\lambda=1.1$ and $\nu=0.25$ ($Q=1.25$) for various lengths N (solid lines). Distributions follow the theoretical result from Eq. **23**. Dashed lines are $p_i=(1-(1-Q)(\alpha+\beta\epsilon_i))^{1/(1-Q)}$ with $\epsilon_i=i-1$. α and β are obtained from fits to the distributions and clearly dependent on N (*Inset*). They can be used to determine c_2.

Assuming that the basis has two relevant components and using the same κ_{ji} as above, the derivative of the constraint term in the example is obtained from Eq. **16**,

$$\frac{d}{dp_i}\log G_{u,T}(pN|\theta) = \log\lambda c_2(pN|\theta)(i-1)p_i^{\nu}. \qquad [25]$$

This constraint term is obviously nonlinear in p_i and is therefore of escort type. Here the expression $\epsilon_i=(i-1)$ plays the role of equidistant energy levels. The example shows explicitly that

finding the most likely distribution function p^* by maximization of $P(k|\theta)$ (minimization of relative entropy) is equivalent to maximizing the generalized entropy of Eq. **24** under a nonlinear constraint term, $\sim\hat{\beta}\left(\sum_i \epsilon_i p_i^Q - U\right)$. In ref. 14 it was shown that a duality exists that allows us to obtain exactly the same result for p^*, when maximizing the dual entropy of Eq. **24**, given by $S^* = (a/Q)\left(\left(1-\sum_{i=1}^W p_i^{2-Q}\right)\Big/(1-Q)\right)$, under the linear constraint, $\beta\left(\sum_i \epsilon_i p_i - U\right)$.

Example: MEP for Path-Dependent Random Processes. We now show that path-dependent stochastic processes exist that are out of equilibrium and whose time-dependent distribution functions can be predicted by the MEP, using the appropriate, system-specific generalized entropy. We consider processes that produce sequences x that increase in length at every step. At a given time the sequence is $x = (x_1, x_2, \ldots, x_N)$. At the next time step a new element x_{N+1} will be added. All elements take one of W different values, $x_i \in \{1, 2, \ldots, W\}$. The system is path dependent, meaning that for a sequence x of length N the probability $p(i|k,\theta)$ for producing $x_{N+1} = i$ depends on the histogram k and the biases θ only. For such processes the probability to find a given histogram, $P(k|\theta)$ can be defined recursively by

$$P(k|\theta) = \sum_{i=1}^{W} p(i|k-e_i,\theta)P(k-e_i|\theta). \qquad [26]$$

For a particular example let the process have the transition probability

$$p(i|k,\theta) = \frac{\theta_i}{Z(k)} \prod_{j=i+1}^{W} g(k_j) \quad \text{with} \quad g(y) = \lambda^{(y^\nu)}, \qquad [27]$$

where $Z(k)$ is a normalization constant, and $\lambda > 0$. Let us further fix $\theta_i = 1/W$. Note that fixing the biases θ in multinomial systems means that as N gets large one obtains $p_i^*(\theta, N) = \theta_i$, for all i. Obviously p^* approaches a steady state and N becomes an irrelevant degree of freedom in the sense that changing N will not change p^*. Fixing all θ_i asymptotically determines p^* completely and leaves no room for any further constraint. For path-dependent processes the situation can be very different. For example, the relative frequencies $p^*(\theta, N)$ of the process defined in Eq. **27** never reach a steady state as N gets larger. [One can show that for such systems the inverse temperature c_2 approximately grows (sub)logarithmically with N.] Here, fixing θ for all i still allows $p^*(\theta, N)$ to evolve with growing N, such that 1 df remains that can be fixed by an additional constraint. [Additional constraints may become necessary for intermediate ranges of N, where some coordinates c_j that need to vanish asymptotically (in the appropriately chosen basis) are not yet sufficiently small.] The process defined in Eq. **27** is a path-dependent, W-dimensional random walk that gets more and more persistent as the sequence gets longer. This means that in the beginning of the process all states are equiprobable ($\theta_i = 1/W$). With every realization of state i in the process, all states $j < i$ become more probable in a self-similar way, and a monotonic distribution function of frequencies emerges as N grows. The process appears to "cool" as it unfolds. Adequate basis vectors $b_{ji}(k)$ can be obtained with deformed multinomials $M_{u,T}(k)$ based on $u(y) = \lambda^{(y^\nu)}$, $T(y) = y$, and a polynomial basis for $\kappa_{ji} = (i-1)^{j-1}$. For this u, in Fig. 1*A* (solid lines), we show normalized deformed binomials for $\nu = 0.7$ and $\lambda = 1.1$. Dashed lines represent the usual binomial. Clearly, generalized multiplicities become more peaked and narrow as N increases, which is a prerequisite for the MEP to hold. In Fig. 1*B* the variance of deformed binomials is seen to diminish as a function of sequence length N for various values of ν. The dashed line shows the variance for the usual binomial. Distribution functions p_i obtained for numerical simulations of sequences with W states are shown in Fig. 1*C* for sequence lengths $N = 1{,}000$, $5{,}000$, and $10{,}000$ (solid lines). Averages are taken over normalized histograms from 150 independent sequences that were generated with $\lambda = 1.1$, and $\nu = 0.25$ ($Q = 1.25$). The distributions follow exactly the theoretical result from Eq. **23**, confirming that a basis with two relevant components (one for the entropy one for a single constraint fixing N) is sufficient for the given process with $\theta_i = 1/W$. Dashed lines are the functions suggested by the theory, $p_i = [1-(1-Q)(\alpha+\beta\epsilon_i)]^{1/(1-Q)}$ with $\epsilon_i = i-1$, where β is obtained from a fit to the empirical distribution. β determines c_2. α is a normalization constant. Although the power exponent $-(1/\nu)$ does not change with N, the "inverse temperature" β increases with N (Fig. 1*C*, *Inset*), which shows that the process becomes more persistent as it evolves—it "ages." Because $T(y) = y$, the observed distribution p can also be obtained by maximizing the generalized entropy S (Eq. **24**) under a nonlinear constraint or, equivalently, by maximizing its dual, S^* with a linear constraint, as discussed above. For other parameter values a basis with more than two components might become necessary. Note that the nonlinear (escort) constraints can be understood as a simple consequence of the fact that the relative frequencies p have to be normalized for all N. In particular, the escort constraints arise from $\sum_i (d/dN)p_i^*(\theta, N) = 0$ and Eq. **23**, which states that p^* does not change its functional shape as θ or N is varied.

Discussion

We have shown that for generalized multinomial processes, where the order of the appearance of events influences the statistics of the outcome (path dependence), it is possible to constructively derive an expression for their multiplicity. We are able to show that a MEP exists for a much wider class of processes and not only for independent multinomial processes. We can explicitly determine the corresponding entropic form from the transition probabilities of a system. We show that the logarithm of the obtained generalized multiplicity is one-to-one related to the concept of Boltzmann entropy. The expressions for the obtained generalized entropies are no longer of Shannon type, $-\sum_i p_i \log p_i$, but assume generalized forms that are known from the entropies of superstatistics (21, 22) and that are compatible with the first three Shannon–Khinchin axioms and violate the fourth (6, 7, 14). Further, we find that generalized entropies are of trace form and are based on known generalized logarithms (14, 16, 18, 23). Our findings enable us to start from a given class of correlated stochastic processes and derive their unique entropy that is needed when using the maximum entropy principle. We are able to determine the time-dependent distribution functions of specific processes, either through minimization of the relative entropy or through maximization of the generalized entropy under nonlinear constraints. A previously discovered duality allows us to obtain the same result by maximization of the dual generalized entropy under linear constraints. Systems for which the new technology applies include out-of-equilibrium, path-dependent processes and possibly even aging systems. In an explicit example of a path-dependent random walk we show how the corresponding generalized entropy is derived. We implement a numerical realization of the process to show that the corresponding maximum entropy principle perfectly predicts the correct distribution functions as the system ages in the sense that it becomes more persistent as it evolves. Systems of this kind often never reach equilibrium as $N \to \infty$.

ACKNOWLEDGMENTS. R.H. and S.T. thank the Santa Fe Institute for hospitality. M.G.-M. acknowledges the generous support of Insight Venture Partners and the Bryan J. and June B. Zwan Foundation.

1. Kittel C (1958) *Elementary Statistical Physics* (Wiley, New York).
2. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27: 379–423, 623–656.
3. Jaynes ET (2003) *Probability Theory: The Logic of Science* (Cambridge Univ Press, Cambridge, UK), pp 351–355.
4. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22(1):79–86.
5. Khinchin AI (1957) *Mathematical Foundations of Information Theory* (Dover, New York).
6. Hanel R, Thurner S (2011) A comprehensive classification of complex statistical systems and an ab-initio derivation of their entropy and distribution functions. *Europhys Lett* 93:20006.
7. Hanel R, Thurner S (2011) When do generalized entropies apply? How phase space volume determines entropy. *Europhys Lett* 96:50003.
8. Bhargava M (2000) The factorial function and generalizations. *Am Math Mon* 107:783–799.
9. Jackson FH (1910) On *q*-definite integrals. *Q J Pure Appl Math* 41(9):193–203.
10. Carlitz L (1938) A class of polynomials. *Trans Am Math Soc* 43:167–182.
11. Polya G (1919) Über ganzwertige polynome in algebraischen zahlkörpern [On integer-valued polynomials in algebraic number fields]. *J Reine Angew Math* 149: 97–116. German.
12. Ostrowski A (1919) Über ganzwertige polynome in algebraischen zahlkörpern [On integer-valued polynomials in algebraic number fields]. *J. Reine Angew. Math.* 149: 117–124. German.
13. Gunji H, McQuillan DL (1970) On a class of ideals in an algebraic numberfield. *J Number Theory* 2:207–222.
14. Hanel R, Thurner S, Gell-Mann M (2011) Generalized entropies and the transformation group of superstatistics. *Proc Natl Acad Sci USA* 108:6390–6394.
15. Beck C, Schlögl F (1995) *Thermodynamics of Chaotic Systems* (Cambridge Univ Press, Cambridge, UK).
16. Hanel R, Thurner S, Gell-Mann M (2012) Generalized entropies and logarithms and their duality relations. *Proc Natl Acad Sci USA* 109(47):19151–19154.
17. Topsoe F (2007) Exponential families and maxent calculations for entropy measures of statistical physics. Complexity, Metastability and Nonextensivity: An International Conference. AIP *Conf Proc* 965:104–113.
18. Hanel R, Thurner S (2007) Generalized Boltzmann factors and the maximum entropy principle: Entropies for complex systems. *Physica A* 380:109–114.
19. Thurner S, Hanel R (2007) Entropies for complex systems: Generalized-generalized entropies. Complexity, Metastability, and Nonextensivity: An International Conference. *AIP Conf Proc* 965:68–75.
20. Tsallis C (1988) A possible generalization of Boltzmann-Gibbs statistics. *J Stat Phys* 52:479–487.
21. Beck C, Cohen EDG (2003) Superstatistics. *Physica A* 322:267–275.
22. Beck C, Cohen EGD, Swinney HL (2005) From time series to superstatistics. *Phys Rev E Stat Nonlin Soft Matter Phys* 72(5 Pt 2):056133.
23. Naudts J (2002) Deformed exponentials and logarithms in generalized thermostatistics. *Physica A* 316:323–334.