Learn your entropy from informative data: An axiom ensuring the consistent identification of generalized entropies

Andrea Somazzi [1,2,*] and Diego Garlaschelli [1,3,4]
[1] IMT School for Advanced Studies, Piazza S. Francesco 19, 55100 Lucca, Italy
[2] Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
[3] Lorentz Institute for Theoretical Physics, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
[4] INdAM-GNAMPA Istituto Nazionale di Alta Matematica, Italy

(Received 15 December 2023; accepted 2 May 2025; published 24 July 2025)

Shannon entropy, a cornerstone of information theory, statistical physics, and inference methods, is uniquely identified by the Shannon-Khinchin or Shore-Johnson axioms. Generalizations of Shannon entropy, motivated by the study of nonextensive or nonergodic systems, relax some of these axioms and lead to entropy families indexed by certain entropic parameters. In general, the selection of these parameters requires preknowledge of the system or encounters inconsistencies. Here we introduce a simple axiom for any entropy family: namely, that no entropic parameter can be inferred from a completely uninformative (uniform) probability distribution. When applied to the Uffink-Jizba-Korbel and Hanel-Thurner entropy families, the axiom selects only Rényi entropy as viable. It also extends consistency with the maximum likelihood principle, which can then be generalized to estimate the entropic parameter purely from data, as we confirm numerically. Remarkably, in a generalized maximum entropy framework the axiom implies that the maximized log-likelihood always equals minus Shannon entropy, even if the inferred probability distribution maximizes a generalized entropy and not Shannon's, solving a series of problems encountered in previous approaches.

DOI: 10.1103/PhysRevResearch.7.033087

I. INTRODUCTION

The concept of entropy was introduced by Clausius in the thermodynamic framework [1] and later adopted in statistical physics by Boltzmann and Gibbs as a tool to describe macroscopic systems in terms of their probabilities of occupancy of microscopic states [2,3]. Within information theory, Shannon axiomatically (re)derived the entropy as a quantification of the uncertainty encoded in a probability distribution, applicable to (among other things) the compressibility of sequences of symbols generated by ergodic probabilistic sources [4]. This allowed Jaynes to subsequently propose that the distribution that maximizes Shannon entropy, under the constraints implied by the empirical information available about a real system and realized via suitable Lagrange multipliers, provides the least biased (maximally noncommittal) inferential description of the unknown microscopic details of that system [5]. For systems whose physical entropy coincides with the Gibbs-Shannon one, this maximum entropy construction can be used to entirely reinterpret statistical physics from an information-theoretic viewpoint. In modern research, statistical inference and model identification based on entropy maximization are perfectly consistent with maximum likelihood estimation methods and are at the heart of several machine-learning techniques [6].

Several generalizations of Shannon entropy (the most popular of which were motivated by the statistical physics of nonextensive and/or nonergodic systems) have been proposed in various contexts [7–11], resulting in extended families of entropy that usually depend, besides the traditional structural parameters (the Lagrange multipliers), on extra entropic parameters that label the specific member of the entropy family. For a fixed choice of these parameters, one can still maximize the resulting entropy and generalize the inference procedure. This is possible when there is enough knowledge *a priori* about the system, so that the entropic parameter can be set by hand to the correct value. However, knowledge about the physical system is often not enough to identify the entropic parameters from first principles. Moreover, the physical entropy and the information-theoretic one may in general no longer coincide for nonextensive or nonergodic systems [12]. Finally, with the available approaches it is generally not possible to maintain compatibility with the maximum likelihood principle and, crucially, to infer the values of the entropic parameters purely from data without encountering inconsistencies, making the generalized methodology inapplicable without prior knowledge of the correct entropy.

In this paper we discuss and alleviate those inconsistencies by introducing an axiom that restricts the form of parametric families of information-theoretic entropies. The axiom enforces a simple uninformativeness requirement and allows for the consistent inference of the entropic parameters purely

*Contact author: andrea.somazzi@imtlucca.it

from the available data, as we show via analytical results and numerical examples. The paper is organized as follows. In Sec. II we first review the theoretical background behind the axiomatic definitions of entropy, the maximum entropy principle, and the maximum likelihood principle. In Sec. III we then discuss the main contributions of the paper, i.e., the introduction of the new axiom and its implications for the selection of information-theoretic entropies from certain popular families, the restoration of consistency with the maximum likelihood principle, and the generalization of the latter in order to infer the entropic parameter(s) purely from the data. Finally, in Sec. IV we offer some concluding remarks.

II. THEORETICAL BACKGROUND

In absence of complete information about a system, one would like to be able to construct an optimal probability distribution providing the least biased inference method (i.e., a probabilistic model) for the unobserved microscopic state of the system, given the partial information available. The maximum entropy principle (MEP) formalizes this idea by first defining an entropy functional as a rigorous quantifier for the degree of uncertainty encoded in a probability distribution, and then looking for the distribution that maximizes this entropy, compatibly with the information available about the system. In this section, we review the various facets of the MEP. First, we illustrate the method given Shannon's definition of entropy, and discuss its connection with the maximum likelihood principle. Then, we recall the assumptions (or axioms) under which Shannon entropy emerges as the unique definition of entropy. Then, we mention examples of attempts that go beyond the traditional axioms and consider generalized definitions of entropy. Finally, we discuss some open challenges that stand in between generalized entropies and the definition of a corresponding generalized MEP, setting the stage for our main results that are discussed in the next section.

A. Maximum entropy principle under Shannon

Given a distribution (technically, a probability mass function) $P = (p(G_1), \dots, p(G_\Omega))$, where $p(G_i)$ is the probability that the discrete random variable G takes the ith outcome (or state) G_i, Ω is the total number of distinct outcomes, and clearly $\sum_{i=1}^{\Omega} p(G_i) = 1$, Shannon entropy $S_1[P]$ is defined as:

$$S_1[P] = -\sum_{i=1}^{\Omega} p(G_i) \ln p(G_i), \tag{1}$$

where the subscript 1 will be justified later. The above expression is unique up to a positive overall multiplicative factor k, which is inessential from the information-theoretic point of view but is important for the identification (when appropriate) with the physical entropy, in which case k carries physical units and coincides with Boltzmann constant. The informational entropy $S_1[P]$ in Eq. (1) coincides (up to Boltzmann's constant) with the physical entropy derived by Gibbs [3], which in turn generalizes Boltzmann entropy [2]. For ergodic and short-range interacting systems, this equivalence is not coincidental and is rooted in statistical inference, as Jaynes showed with the introduction of the maximum entropy principle (MEP) [5]. The MEP states that, given only a set I of pieces of empirical information about a system (in the physical situation, this typically means the knowledge of a few, macroscopic conserved quantities such as the total energy and/or the total number of particles), one should assign the possible microscopic states a probability distribution P that maximizes the entropy. In other words, entropy can be used as an inference functional whose maximization minimizes bias and prevents arbitrariness.

In particular, consider a system with a set of Ω potential microstates $\{G_i\}_{i=1}^{\Omega}$ and assume that the available information I is encoded in the empirical value $C^* = C(G^*)$ of a certain (scalar or vector) function C of the microstate of the system, where G^* is the true (unobservable) microstate. For the moment, let us assume that C^* is the only observation available (later, we will consider multiple observations of the same variable). Since G^* is unknown, the microstate is treated as a random variable G. The MEP applied to $S_1[P]$ identifies the maximum-entropy distribution for G, which we denote as $P_0 = (p_0(G_1), \dots, p_0(G_\Omega))$ or $P_1 = (p_1(G_1), \dots, p_1(G_\Omega))$, depending on whether C^* is treated as a hard or soft constraint, respectively.

In the case of hard constraints (microcanonical ensemble), only a restricted number $\Omega_{C^*} < \Omega$ of microstates i for which $C(G_i)$ matches C^* exactly are assigned a nonzero probability, which has to be uniform over the restricted support, i.e., $p_0(G_i) = \Omega_{C^*}^{-1}$ if $C(G_i) = C^*$ and $p_0(G_i) = 0$ otherwise. The resulting entropy is

$$S_1[P_0] = \ln \Omega_{C^*}. \tag{2}$$

Unfortunately, calculating Ω_{C^*} is generally a hard combinatorial problem, which makes the microcanonical ensemble not amenable to analytical calculations. For this reason, soft constraints are considered more often in the literature, as we also do in this paper.

In the case of soft constraints (canonical ensemble), only the expected value $\langle C \rangle$ of the observable is constrained to match C^*, i.e.,

$$\langle C \rangle \equiv \sum_{i=1}^{\Omega} p(G_i) C(G_i) = C^*, \tag{3}$$

thus allowing for the full set of Ω microstates, however, with a nonuniform probability $p(G_i)$ yet to be determined. To find the specific probability $p_1(G_i)$ maximizing S_1 under the soft constraint above, one can introduce the Lagrange multiplier θ (which has the same dimensionality as C), plus an additional scalar multiplier α enforcing the normalization of P, and look for the specific values (denoted as P_1, θ_1, α_1) corresponding to the stationary point for which all the derivatives of the Lagrangian function

$$\mathcal{L}_1[P] \equiv S_1[P] - \alpha \left[\sum_{i=1}^{\Omega} p(G_i) - 1 \right] - \theta \cdot [\langle C \rangle - C^*] \tag{4}$$

vanish (the notation $\theta \cdot C$ indicates the scalar product). Setting the different derivatives (with respect to P_1, α_1 and θ_1 respectively) to zero produces different aspects of the final result. Setting $\partial \mathcal{L}[P]/\partial P|_{P_1} = 0$, i.e., $\partial \mathcal{L}[P]/\partial p(G_i)|_{p_1(G_i)} = 0 \; \forall i$, (first vanishing component of the derivative) leads to the

functional form of P_1, which turns out to be the well-known Boltzmann-Gibbs distribution

$$p_1(G_i,\theta)=\frac{e^{-\theta\cdot C(G_i)}}{Z_1(\theta)},\quad Z_1(\theta)=\sum_{j=1}^{\Omega}e^{-\theta\cdot C(G_j)}, \tag{5}$$

where $Z_1(\theta)$ is the partition function, resulting from the normalization constraint (second vanishing component of the derivative)

$$\left.\frac{\partial\mathcal{L}[P_1]}{\partial\alpha}\right|_{\alpha_1}=0\Rightarrow\sum_{i=1}^{\Omega}p_1(G_i,\theta)=1, \tag{6}$$

which leads to

$$\alpha_1=-1+\ln Z_1(\theta) \tag{7}$$

independently of the value of θ. Importantly, P_1 is not identified entirely, until the parameter θ is also determined. This is attained by enforcing the vanishing of the third component of the derivative, which identifies the value θ_1 realizing Eq. (3):

$$\left.\frac{\partial\mathcal{L}[P_1]}{\partial\theta}\right|_{\theta_1}=0\Rightarrow\sum_{i=1}^{\Omega}p_1(G_i,\theta_1)C(G_i)=C^*, \tag{8}$$

where, if θ is a vector, the notation means again that all the derivatives of $\mathcal{L}[P]$ with respect to the components of θ vanish separately. The final solution to the MEP problem is therefore given by inserting θ_1 into Eq. (5), and we will denote it as $P_1(\theta_1)=(p_1(G_1,\theta_1),\ldots,p_1(G_\Omega,\theta_1))$. The MEP with soft constraints, which are appropriate when the observables are expected to fluctuate, has been used successfully for inference and model selection in many fields beyond physics, including network theory, neuroscience, economics and biology [13,14].

B. Maximum likelihood principle

It is very important to realize that the MEP procedure outlined above has deep connections and desirable consistencies with the maximum likelihood (ML) principle, which applies to more general (not necessarily maximum-entropy) parametric probability distributions and states that the optimal parameter value θ^* is the one maximizing the log-likelihood on the data G^*. If applied to the exponential family from Eq. (5), obtained by setting to zero only the first and second components of the derivative of the Lagrangian, the ML principle would select the value

$$\theta_1^*=\text{argmax}_\theta\ell_1(\theta),\quad \ell_1(\theta)\equiv\ln p_1(G^*,\theta). \tag{9}$$

As a first result, it is easy to show that the value θ_1^* defined by Eq. (9) coincides with the value θ_1 that would be obtained by setting also the third component of the derivative to zero as in Eq. (8) [15], i.e., $\theta_1^*\equiv\theta_1$ (in our notation, the asterisk next to a parameter will always denote the ML value of that parameter), i.e.,

$$\left.\frac{\partial\ell_1(\theta)}{\partial\theta}\right|_{\theta_1^*}=0\Rightarrow\sum_{i=1}^{\Omega}p_1(G_i,\theta_1^*)C(G_i)=C^*, \tag{10}$$

which indeed coincides with the result in Eq. (8). This means that the ML principle can be seen as equivalent to the part of the Lagrangian optimization relative to θ.

Moreover, it is straightforward to show that the maximized log-likelihood equals minus the entropy:

$$S_1[P_1(\theta_1^*)]=-\ell_1(\theta_1^*), \tag{11}$$

which is the counterpart of Eq. (2) in the case of soft constraints. This relationship is very important, because the maximized likelihood is at the basis of model selection criteria [16,17]: if alternative models (i.e., alternative parametric probability distributions) are compared against the same empirical data, the model to be preferred (assuming all models have the same complexity, e.g., the same number of parameters) is the one with highest maximized likelihood. Then, Eq. (11) ensures that the ranking of models based on ML is the same as the ranking based on minus their entropy: the least uncertain (i.e., most informative) model has to be preferred. For models with different numbers of parameters and/or functional forms, the ranking based on likelihood/entropy has to be revised, usually by removing from the maximized log-likelihood a term quantifying the model complexity, leading to criteria such as Akaike's information criterion (AIC), the Bayesian information criterion (BIC), the minimum description length (MDL), etc. [16,17]. While we will not consider this situation here for simplicity, we note that Eq. (11) remains perfectly consistent with the above criteria: in particular, the key idea behind MDL is that the total description length of the data, given a model, is calculated as the minimum length (after maximal compression) of the data given the model (which is exactly given by the entropy $S_1[P_1(\theta_1^*)]$), plus the entropy of the model itself (which is the part, not considered here, that is larger for more complex models) [17]. Also note that when the maximum-entropy distribution $P_1(\theta_1^*)$ is inserted into Eq. (4), we get

$$\mathcal{L}_1[P_1(\theta_1^*)]=S_1[P_1(\theta_1^*)]=-\ell_1(\theta_1^*), \tag{12}$$

from which we learn that the Lagrangian, evaluated at its stationary point $P_1(\theta_1^*)$, coincides with minus the maximized log-likelihood and can therefore be used to rank alternative models as well. All the above results indicate that the MEP can be used as a model selection criterion, exactly as the ML principle, by ranking models of equal complexity based uniquely on their realized entropy.

It is also important to notice that it is actually the ML principle that produces the interpretation for the value C^* appearing in Eq. (3). Indeed, while we started our discussion defining C^* as $C(G^*)$, where G^* is the unknown microstate of the system, there is actually nothing in Eq. (3), and consequently in Eq. (8) (which follows from the MEP), that guarantees that the (natural) definition of C^* is in fact $C(G^*)$. In other words, the only interpretation we get for C^* from the MEP is that it is the target value for the ensemble average $\langle C\rangle$. By contrast, Eq. (8) (which follows from the ML principle) does indicate that $C^*=C(G^*)$. In this sense, the ML principle is more informative than the MEP in identifying what is the definition (as a function of G^*) one should choose for C^* as a target value for $\langle C\rangle$.

To explain this point more clearly, let us consider the case when there are M independent observations $\{C_m^*\}_{m=1}^M$ about the system, which technically means that there are M independent and identically distributed (i.i.d.) realizations $\{G_m^*\}_{m=1}^M$ of the microstate G (recall that G is treated as a random variable),

on each of which the quantity $C_m^* = C(G_m^*)$ $(m = 1, \cdots, M)$ is observed. Clearly, since the system being observed multiple times is one, the probability distribution characterizing it must still be specified by a single value of the Lagrange multiplier θ coupled to the quantity C. It should at this point be clear that the principle that identifies how to optimally combine the M observations $\{C_m^*\}_{m=1}^M$ in order to estimate θ is not the MEP, but the ML one. Indeed, the ML principle applied to the joint log-likelihood $\sum_{m=1}^M \ln p_1(G_m^*, \theta)$, or equivalently to the average log-likelihood $\bar{\ell}_1(\theta) \equiv \sum_{m=1}^M \ln p_1(G_m^*, \theta)/M$, can be formulated by replacing Eq. (9) with

$$\theta_1^* = \text{argmax}_\theta \bar{\ell}_1(\theta), \quad \bar{\ell}_1(\theta) \equiv \frac{\sum_{m=1}^M \ln p_1(G_m^*, \theta)}{M}. \quad (13)$$

It is easy to show that the condition $\partial\bar{\ell}_1(\theta)/\partial\theta|_{\theta_1^*} = 0$ identifying θ_1^* leads to the well-known result

$$\langle C \rangle = \frac{1}{M}\sum_{m=1}^M C_m^*, \quad (14)$$

where the (arithmetic) sample average of the M observations has emerged. So, in order to find the ML parameter value θ_1^*, one should replace Eq. (3) with Eq. (14), or equivalently redefine C^* in Eq. (3) as the sample average of $\{C_m^*\}_{m=1}^M$. In plain words, the sample average is produced by the ML principle. On the contrary, within the MEP construction, there is no way of telling Eqs. (4) and (8) what, in case of M observations, the meaning and definition of C^* should be. So, again, the ML principle is in this sense more informative than the MEP. This is another reason why one wants the entropy to be fully consistent with what the ML principle leads to. In particular, it is easy to show that, due to the assumed independence of the M samples, the maximized average log-likelihood $\bar{\ell}_1(\theta_1^*)$ still equals minus the entropy:

$$S_1[P_1(\theta_1^*)] = -\bar{\ell}_1(\theta_1^*), \quad (15)$$

generalizing Eq. (11). Note that there is no microcanonical counterpart of Eq. (15), since Eq. (2) cannot be generalized to the case $M > 1$, unless all the M values $\{C_m^*\}_{m=1}^M$ are identical. Indeed the microcanonical ensemble cannot be constructed, because by definition it cannot account for different realizations of the values of the constraints: in case of different observed values of the same constraint, only the canonical ensemble is feasible.

The above discussion clarifies that it is important that the entropy is consistent with the maximized log-likelihood, because the ML principle is needed both for model selection and for the determination of how multiple observations of the same system should be combined in order to optimally estimate the parameters.

C. Shannon-Khinchin axioms

As anticipated, we now come to the origin of the expression for the entropy in Eq. (1). From an information-theoretical point of view, Shannon entropy is axiomatically defined through the following four Shannon-Khinchin (SK) axioms [18]:

(1) *SK1 (continuity):* $S[P]$ is continuous in the entries of P.

(2) *SK2 (maximality):* $S[P]$ is maximal when P is the uniform distribution $P_u \equiv (\Omega^{-1}, \ldots, \Omega^{-1})$.

(3) *SK3 (expansibility):* $S[P]$ is expansible, i.e., it does not change if for the variable G an $(\Omega + 1)$-th outcome with zero probability $(p(G_{\Omega+1}) = 0)$ is added:

$$S[(p(G_1), \ldots, p(G_\Omega))] = S[(p(G_1), \ldots, p(G_\Omega), 0)].$$

(4) *SK4 (separability):* the entropy of the joint distribution $R = (r(G_1^{(1)}, G_1^{(2)}), \ldots, r(G_{\Omega^{(1)}}^{(1)}, G_{\Omega^{(2)}}^{(2)}))$ of two variables $G^{(1)}$ and $G^{(2)}$ with marginal distributions $P = (p(G_1^{(1)}), \ldots, p(G_{\Omega^{(1)}}^{(1)}))$ and $Q = (q(G_1^{(2)}), \ldots, q(G_{\Omega^{(2)}}^{(2)}))$, respectively, where $p(G_i^{(1)}) = \sum_{j=1}^{\Omega^{(2)}} r(G_i^{(1)}, G_j^{(2)})$ and $q(G_j^{(2)}) = \sum_{i=1}^{\Omega^{(1)}} r(G_i^{(1)}, G_j^{(2)})$, separates as

$$S[R] = S[P] + S[Q|P].$$

Here $S[Q|P]$ is the conditional entropy of Q on P, defined as $S[Q|P] = \sum_{k=1}^{\Omega^{(1)}} p(G_k^{(1)}) S[Q_{|k}]$ with $Q_{|k} = (r(G_k^{(1)}, G_1^{(2)})/p(G_k^{(1)}), \ldots, r(G_k^{(1)}, G_{\Omega^{(2)}}^{(2)})/p(G_k^{(1)}))$ denoting the conditional distribution of the events in Q on the kth event in P. Note that in particular, if the two events are independent ($Q = Q_{|k}$ for all k), then $S[R] = S[P] + S[Q]$, in which case separability becomes additivity.

It is possible to show that the only functional form of $S[P]$ respecting the four SK axioms is Shannon entropy $S_1[P]$. Also, as required by *SK2*, the maximum value of $S_1[P]$ is attained by the uniform distribution P_u, leading to Boltzmann entropy:

$$S_1[P_u] = \ln \Omega. \quad (16)$$

No distribution P can be such that $S_1[P] > S_1[P_u]$.

While these axioms uniquely lead to Shannon entropy, there exists relaxations (in particular of *SK4*) leading to generalized entropies, e.g., Rényi or Tsallis entropy, which we will discuss later in this paper.

D. Shore-Johnson axioms

An alternative axiomatic definition of entropy was proposed by Shore and Johnson (*SJ*) in terms of requirements imposed not directly on the entropy functional, but on the resulting probability distribution, denoted as $P = \circ I$, obtained by maximizing the entropy in presence of a set I of pieces of information [19]. The *SJ* axioms are the following:

(1) *SJ1 (uniqueness):* given I, $P = \circ I$ is unique.

(2) *SJ2 (invariance):* if $\Gamma[\cdot]$ is a coordinate transformation (change of variables), then $\Gamma[\circ I] = \circ(\Gamma[I])$.

(3) *SJ3 (system independence):* given two independent systems A and B, it should not matter whether one accounts for distinct pieces of information about them separately (in terms of marginal probabilities) or jointly (in terms of a joint probability). This means $\circ(I_A \wedge I_B) = (\circ I_A)(\circ I_B)$, where $I_A \wedge I_B$ denotes the union of the available pieces of information I_A and I_B about A and B respectively.

(4) *SJ4 (subset independence):* it should not matter whether one treats an independent subset of system states in terms of a separate conditional density or in terms of the full system density. Consider a partition of the system's states into disjoint subsets $\{\Lambda_k\}_k$ such that $\bigcup_k \Lambda_k = \Omega$, for each k of which there is a piece of information I_k available. Then

$(\circ I)_{\Lambda_k} = \circ I_k \; \forall k$, where $I = \bigwedge_k I_k$ is the total information, and $P_{\Lambda_k} = (p_{\Lambda_k}(G_1), \ldots, p_{\Lambda_k}(G_\Omega))$, where $p_{\Lambda_k}(G_i) = p(G_i | G_i \in \Lambda_k)$ denotes the conditional distribution relative to the subset Λ_k.

(5) *SJ5 (maximality):*[1] with no information available ($I = \varnothing$), $P = \circ I$ is the uniform distribution P_u.

Shore and Johnson claimed that Shannon entropy is the only inference functional compatible with their axioms, a statement suggesting the equivalence of the SK and the SJ axioms. However, it was later clarified [20,21] that Shore and Johnson's conclusion was due to an additional hidden assumption they made inadvertently when formally using *SJ3* in their reasoning. Specifically, they considered a situation where distinct pieces of information I_A and I_B are known about two systems A and B, and implied that the resulting joint probability factorises as $\circ(I_A \wedge I_B) = (\circ I_A)(\circ I_B)$, thereby applying *SJ3* even if the independence of the two systems is not guaranteed (having only disjoint pieces of information about two systems does not guarantee that the two systems are independent) [20,21]. The presence of this additional assumption implies that Shannon entropy is in fact the desired functional only when systems are independent: if this is not the case, then the resulting maximum entropy distribution is no longer maximally noncommittal with respect to missing information, as Jaynes' MEP demands it to be [5], because there is actually no information available about the (in)dependence of the systems.

E. Generalized entropies

Uffink [20] showed that, if Shore and Johnson's proof is correctly revisited without the extra unjustified assumption, the entropy resulting from the SJ axioms is not uniquely determined and is actually an entire generalized family $S_q^{(f)}[P]$, given by any increasing function f of a certain functional $U_q[P]$ that we will call the Uffink functional, i.e.,

$$S_q^{(f)}[P] = f(U_q[P]), \quad U_q[P] = \left(\sum_{i=1}^{\Omega} p^q(G_i) \right)^{\frac{1}{1-q}} \tag{17}$$

for some parameter $q > 0$. For a given f, each entropy in the family is identified by the parameter q, which we will therefore call the entropic parameter. Note that an entropic parameter plays a different role with respect to other structural parameters entering the entropy, such as θ in the Shannon case discussed above. Clearly, Shannon entropy is one of the possible members of this family. Indeed, taking $f(x) = \ln x$, one can show that

$$\lim_{q\to 1} S_q^{(\ln)}[P] = \lim_{q\to 1} \ln U_q[P] = -\sum_{i=1}^{\Omega} p(G_i) \ln p(G_i) = S_1[P]. \tag{18}$$

In other words, Shannon entropy formally corresponds to $q = 1$, justifying the subscript adopted in Eq. (1) (note instead that the subscript in the uniform distribution P_0 used in Sec. II A to describe the microcanonical distribution under hard constraints has nothing to do with the case $q = 0$, which is inadmissible). Notably, Jizba and Korbel [21,22] showed that an entropy of the type $f(U_q[P])$ can also be obtained from the SK axioms, provided that *SK4* is relaxed to a generalized separability condition where the sum is replaced by the so-called Kolmogorov-Nagumo sum,[2] previously introduced in the context of generalized arithmetics [23,24]. This shows that the SJ axioms are actually equivalent to a specific generalization of the SK ones. The generalization of *SK4* has been a matter of discussion in the statistical physics literature for decades, as it relates to the subject of nonextensive (or rather nonadditive) thermodynamics [7]. We will call any entropy of the form $f(U_q[P])$ an Uffink-Jizba-Korbel (UJK) entropy.

Several other generalized families of entropy resulting from relaxations of the SK or SJ axioms have been proposed [9,10]. A notable example is the so-called (c, d) entropies $S_{c,d}[P]$ introduced by Hanel and Thurner [11,25] by replacing *SK4* with the assumption of trace-form (or more in general composable) entropies, i.e., entropies that can be written as (functions of) a sum over the states $\{G_i\}_{i=1}^{\Omega}$ of the system. In particular, an entropy $S(P)$ is trace form if it can be written as a sum $\sum_{i=1}^{\Omega} g(p(G_i))$ for some function g. Note that Shannon entropy is in this class, with $g(x) = -x \ln x$. More generally, a composable entropy can be written as a function h of such a sum, i.e.

$$S_{c,d}^{(h,g)}[P] = h\left(\sum_{i=1}^{\Omega} g(p(G_i)) \right), \tag{19}$$

where the entropic parameters (c, d) are determined by how the entropy scales with the number Ω of accessible configurations [8,25]. In particular, one considers the transformations $\Omega \to \lambda\Omega$, $\Omega \to \Omega^{1+a}$ and identifies c and d from the following limiting ratios:

$$\lim_{\Omega\to\infty} \frac{S_{c,d}^{(h,g)}[(p(G_1), \ldots, p(G_{\lambda\Omega}))]}{S_{c,d}^{(h,g)}[(p(G_1), \ldots, p(G_\Omega))]} = \lambda^{1-c}, \tag{20}$$

$$\lim_{\Omega\to\infty} \frac{S_{c,d}^{(h,g)}[(p(G_1), \ldots, p(G_{\Omega^{1+a}})]}{S_{c,d}^{(h,g)}[(p(G_1), \ldots, p(G_\Omega))]} \Omega^{a(c-1)} = (1+a)^d. \tag{21}$$

Different choices of h and g may result in the same values of the entropic parameters, in which case the corresponding entropies are considered asymptotically equivalent [8]. Therefore in this case the entropic parameters identify equivalence classes of entropies with the same asymptotic properties. We

[1] Actually, Shore and Johnson defined the maximality axiom only implicitly. Indeed, starting from the principle of minimum cross-entropy, they introduced the MEP as its equivalent in the case where the prior distribution is uniform. For this reason, even if not explicitly axiomatized, they considered the posterior P to be equal to the uniform distribution (i.e., the same as the prior) when no information is available.

[2] Considering a bijection $f^{-1} : M \mapsto N \subset \mathbb{R}$, the generalized arithmetics is defined as follows:

$$x \oplus y = f(f^{-1}(x) + f^{-1}(y)),$$
$$x \ominus y = f(f^{-1}(x) - f^{-1}(y)),$$
$$x \otimes y = f(f^{-1}(x) f^{-1}(y)),$$
$$x \oslash y = f(f^{-1}(x) / f^{-1}(y)).$$

will call the entropies that respect *SK1–SK3*, plus Eq. (19), the Hanel-Thurner (HT) entropies.

F. How to identify the correct entropy

On UJK entropies, HT entropies and in principle any generalized entropy family, it is important to ensure that the MEP can be reformulated consistently as a tool to construct probability distributions starting from observations of the system. This procedure is sometimes called the generalized maximum entropy principle (GMEP). However, two conceptual and practical problems are currently open.

First, while it is still possible, for a fixed value of the entropic parameter(s), to identify the functional form of the probability distribution maximizing the generalized entropy under certain soft constraints, it is no longer guaranteed in general that the enforcement of these constraints remains consistent with the application of the ML principle to the Lagrange multipliers and that the entropy retains a role for model selection as in Eq. (11). Only for certain generalized entropies this consistency is retrieved, but not for all of them, as we show later with some notable examples. Since the ML principle is agnostic with regard to the form of the probability distribution, and even more so to the type of entropy the latter maximizes, this inconsistency raises suspicion. Unfortunately, its possible origin is poorly discussed in the literature.

Second, fundamental problems arise when considering the determination of the entropic parameters themselves. In particular, two main approaches have been proposed. One approach requires some *a priori* knowledge of the system (e.g., how certain properties of the entropy or of the system change with the number of accessible configurations [8,25,26]) as in Eqs. (20) and (21), implying that, in absence of such knowledge, the entropic parameters cannot be consistently derived purely from data as the other parameters (in this approach, the knowledge of the entropic parameter is viewed as a different type of information, besides the information obtained by the empirical measurement). Another approach does allow for the entropic parameters to be inferred from data, again invoking some form of maximization of the generalized entropy [27,28]. However, as we show below, this requirement conflicts with the ML principle, if the latter is extended to the estimation of the entropic parameters themselves. Finally, several analyses estimate the entropic parameter(s) by assuming a certain form of the entropy and fitting the resulting maximum-entropy probability on empirical distributions [29–42]. However, we would like to stress that maximum-entropy distributions, even when optimally fitted to the data, do not uniquely identify the entropy they maximize, because they also maximize any other entropy functional that is a monotonic function of that entropy. This point becomes particularly critical when such monotonic function depends on the entropic parameter(s) themselves, as we will show below.

In the following section we introduce an axiom that, given an entropic family, selects a specific entropic function. Subsequently, we apply ML to the resulting maximizing probability distribution to estimate the entropic parameter(s). The combination of both the proposed axiom and the ML principle enables the comprehensive determination of the entropy.

III. THE UNINFORMATIVENESS AXIOM

The limitations discussed in the previous section make the GMEP either inapplicable in practice without prior knowledge of the correct entropic parameter(s), or inconsistent with the ML principle and the information-theoretic consequences of *SJ3* under independence. In the rest of this section, which contains our main results, we show that a possible solution to this problem can be achieved starting from a seemingly different viewpoint, i.e., by imposing an additional axiom that somehow aligns all entropies in a given family and therefore allows us to select the most likely member of the family purely from data (if the latter contain information) and without prior knowledge of the system's properties. Remarkably, the introduction of this simple requirement solves all the inconsistencies discussed in Sec. II F.

A. Uninformativeness axiom

We now introduce the axiom. Unlike the SK or SJ ones, this axiom applies not to an individual entropy in a generalized parametric family, but rather to the entire family. Indeed the axiom does not represent yet another generalization of the SK or SJ ones, but rather an auxiliary requirement to be added precisely when any such generalization is made, to restrict the form of the resulting entropic family.

(1) *Uninformativeness axiom*: In a parametric family of information-theoretic entropies, the value of the entropy attained by the uniform distribution P_u should not depend on the value of the entropic parameter(s).

Clearly, if the axiom is applied to families that include Shannon entropy S_1 as a particular case, it implies that all members of the family attain the same value $S_1[P_u] = \ln \Omega$ when applied to P_u. This requirement equips generalized entropies with a universal scale and meaning. As we show below, our axiom provides certain guarantees when the inference procedure is extended to the identification of the entropic parameters themselves. On one hand, the axiom ensures that no entropic parameter can be inferred from a completely uninformative (i.e., uniform) distribution, irrespective of how the parameter estimation procedure is conceived. On the other hand, when informative (nonuniform) data are available, the axiom ensures consistency with a generalized ML principle and model selection approach where all parameters, including the entropic one, can be identified from empirical observations, without prior knowledge of the system. Note that, in general, the physical entropy characterizing the real system may be different from the information-theoretic one identified by our axiom; nonetheless, our axiom ensures that the maximum-entropy distribution that best describes the physical observations can be identified consistently from the information-theoretic entropy, without prior knowledge of the physical entropy itself.

Note that, as required by *SK2* and *SJ5*, for a given value of q the Uffink functional in Eq. (17) is maximized by P_u. This requirement comes from a horizontal perspective, in the sense that it holds for each q entropy in the family. Our axiom, on

the other hand, provides a vertical perspective: among all the q entropies, none of them has to be preferred when applied to P_u. In other words, the axiom ensures the uninformativeness role of the uniform distribution not only for a specific entropy in the family, but across all of them. Since *SK2* and *SJ5* ensure that no entropy can exceed the value it attains on P_u, the axiom establishes a sort of common reference frame or universal scale, which allows us to compare different entropies in a parametric family consistently. In particular, it ensures that all entropies in a parametric family that respects *SK2* or *SJ5* and includes Shannon entropy as a particular case attain values in the same interval $[0, \ln \Omega]$, irrespective of the value of the entropic parameter(s). We will show that this guarantee ensures that the entropic parameter(s) can be estimated via a model selection approach purely from the input data, if the latter are informative (nonuniformly distributed).

B. Application to important entropy families

We now discuss some consequences of imposing the uninformativeness axiom to popular entropy families.

We start with the UJK entropies $S_q^{(f)}[P]$ under the requirement that the family should include Shannon entropy as a particular case. The entropy $S_q^{(f)}[P] = f(U_q[P])$, when evaluated on the uniform probability distribution $P_u = (\Omega^{-1}, \ldots, \Omega^{-1})$, returns the value

$$S_q^{(f)}[P_u] = f(U_q[P_u]) = f(\Omega) \quad \text{for} \quad q \neq 1. \tag{22}$$

Our axiom requires that $S_q^{(f)}[P_u]$ is independent of q, which implies that f should be independent of q. For $q = 1$, technically $S_q^{(f)}[P]$ is only defined as the limit

$$\lim_{q\to 1} S_q^{(f)}[P] = f(\lim_{q\to 1} U_q[P]), \tag{23}$$

where we have used the q independence of f. If we require that, when $P = P_u$, this limit coincides with what Shannon entropy returns on P_u, i.e., $S_1[P_u] = \ln \Omega$, then we need a function f such that

$$\lim_{q\to 1} S_q^{(f)}[P_u] = f(\lim_{q\to 1} U_q[P_u]) = \ln \Omega, \tag{24}$$

i.e., $f(x) = \ln x$. Therefore, combining Eqs. (22) and (24) we obtain $f(x) = \ln x$ for all q, i.e., the only viable UJK entropy is Rényi entropy [43]

$$S_q[P] \equiv S_q^{(\ln)}[P] = \ln U_q[P] = \frac{1}{1-q} \ln \sum_{i=1}^{\Omega} p^q(G_i), \tag{25}$$

where, since the entropy above is the only surviving one in the family $S_q^{(f)}[P]$, we have removed the superscript from the resulting $S_q^{(\ln)}[P]$. From Eq. (18) we can confirm that this entropy reduces to Shannon entropy in the limit $q \to 1$, a well-known result for Rényi entropy. This entropy is such that, on the uniform distribution P_u,

$$S_q[P_u] = \ln \Omega, \tag{26}$$

which does not depend on q, as demanded by our axiom. Therefore the only viable UJK entropy is Rényi entropy. In general, other UJK entropies do not respect our axiom.

An important counterexample is Tsallis entropy [44], defined as

$$S_q^{\text{Tsallis}}[P] \equiv S_q^{(\ln_q)}[P] = \frac{1}{1-q}\left(\sum_{i=1}^{\Omega} p^q(G_i) - 1\right) \tag{27}$$

and obtained from the so-called q logarithm $f(x) = \ln_q(x) \equiv (x^{1-q} - 1)/(1-q)$ (not to be confused with the ordinary logarithm of x to base q): indeed, when evaluated on P_u, this entropy takes the q-dependent value

$$S_q^{\text{Tsallis}}[P_u] = \frac{\Omega^{1-q} - 1}{1-q} = \ln_q(\Omega). \tag{28}$$

From the point of view of our axiom, such q dependence is a contradiction: different values of q should not artificially attach different degrees of informativeness to an intrinsically uninformative distribution. Seen from another point of view, this contradiction arises from the q dependence of the function f defining Tsallis entropy from the Uffink functional $U_q[P]$: such q dependence is not admitted by our axiom because $f(U_q[P_u])$ should not depend on q. Note that the q independence of the function f defining the UJK entropy $f(U_q[P_u])$ is a nontrivial consequence of our axiom, as it arises as necessary only when comparing entropies obtained for different values of q (if only a single value of q were considered, nothing would prevent f from being specified by that value of q). In particular, our axiom would demand $q = 1$ in order to have $S_q^{\text{Tsallis}}[P_u] = S_1[P_u]$, i.e., the only viable Tsallis entropy is Shannon entropy. We should stress at this point that the inadmissibility of Tsallis entropy under our axiom is not in contradiction with the many successful empirical applications of the so-called q-exponential or Tsallis distribution maximizing Tsallis entropy for fixed q [7,29–42], because such distribution (that we explicitly consider later in this paper) is exactly the same as the one maximizing Rényi entropy or any other monotonic function of the Uffink functional, as we also discuss below. However, when that distribution is put back into the entropy, only Rényi entropy gives consistent results in terms of the absolute quantification of the uncertainty and the associated ML estimation and model selection procedures. Indeed, we will show that a ranking of models (or values of q) based on Tsallis entropy would mess up the ranking based on ML, while the use of Rényi entropy restores and extends the consistency with the ML principle.

As another example, we apply the uninformativeness axiom to the HT family of composable (c, d) entropies that can be written as in Eq. (19). If we require $S_{c,d}^{(h,g)}[P_u] = S_1[P_u] = \ln \Omega$ in analogy with Eq. (26), then the axiom translates Eqs. (20) and (21) to:

$$\lim_{\Omega\to\infty} \frac{\ln \lambda\Omega}{\ln \Omega} = \lambda^{1-c} \tag{29}$$

$$\lim_{\Omega\to\infty} \frac{\ln \Omega^{1+a}}{\ln \Omega} = (1+a)^d \tag{30}$$

and implies $(c, d) = (1, 1)$. This parameter choice identifies the equivalence class of entropies that are additive for independent events. Both Shannon and Rényi entropies belong to this class. In particular, in the case $h(x) = x$ (trace-form entropy) and $g(x) = -x \ln x$ (Shannon entropy), one gets $(c, d) = (1, 1)$ [8], i.e., $S_{1,1}^{(x,-x\ln x)}[P] = S_1[P]$. Therefore

Shannon entropy is a viable trace-form HT entropy under our axiom. Similarly, in the case $h(x) = \ln(x)/(1-q)$ and $g(x) = x^q$ (Rényi entropy) one again gets $(c, d) = (1, 1)$ [8], i.e., $S_{1,1}^{(\ln(x)/(1-q),x^q)}[P] = S_q[P]$. Therefore Rényi entropy is a viable composable HT entropy. By contrast, the case $h(x) = x$ and $g(x) = (x^q - \Omega^{-1})/(1-q)$ (Tsallis entropy) leads to $(c, d) = (q, 0)$ [8], confirming that Tsallis entropy (which is another trace-form entropy) does not respect our axiom.

The fact that, for both the UJK and HT families, only Rényi entropy (or an asymptotically equivalent one) survives our axiom does not disagree with the possibility of nonextensivity of the entropy, which has led to the introduction of many variants of entropy over the last decades [7,8]. Indeed, while our axiom selects entropy additivity for independent systems (as both Shannon and Rényi do), it does not have direct implications when independence is not present or even not known. In particular, it should be stressed that nonextensivity is a property not of the entropy itself, but of how the number Ω of configurations scales with the physical size of the system (i.e., the number n of units or particles) [8]. Even Shannon entropy can be nonadditive if applied to a system where Ω (or Ω_{C^*}, when in presence of a constraint C^*) is not exponential in n, as clear from Eq. (16) or (2). Note that Eq. (2) applies in the microcanonical case, but a similar nonextensive scaling of the entropy would be exhibited in the canonical case as well, where $e^{S_1[P_1(\theta_1^*)]}$ is the counterpart of Ω_{C^*} and plays the role of the effective number of (typical) realizations. An important example in this respect is provided by random graphs: the number of all binary graphs on n vertices is $\Omega = 2^{\binom{n}{2}}$, so it is superexponential [8,45]. Even when subject to various types of constraints C^*, the (effective) number Ω_{C^*} remains superexponential [14]. At the opposite extreme, even for systems where Ω does increase exponentially in n, the system may still be subject to certain constraints such that Ω_{C^*} is subexponential in n, so that the resulting entropy is subextensive. An example is the class of state space reducing processes [25]. Later in the paper, we will show that Shannon entropy can grow nonlinearly in n even for a simple example of n independent observations. Therefore one first general result implied by the uninformativeness axiom is that nonextensivity or nonergodicity (when present) should be completely encoded in the scaling of Ω_{C^*} with n, thus ultimately in the identification of the proper (effective) constraint C^*, and not in the expression of the entropy itself. Finally, we notice that, while for the considered UJK and HT entropies the uninformativeness axiom leads to the same result that would be obtained by enforcing the requirement of additivity of entropy for independent processes, this does not imply that the latter requirement is in general equivalent to the axiom itself: indeed, the axiom can in principle be applied to any parametric entropy family, for which different results, irreducible to the requirement of additivity, might follow. Future research could explore whether generalized entropic families exist for which the uninformativeness axiom and the additivity requirement would lead to different results.

C. Generalized MEP

In a GMEP context, a direct consequence of the fact that our axiom restricts the viable expressions for the generalized entropies is, of course, a corresponding restriction on the probability distributions maximizing such generalized entropies under soft constraints (note that, under hard constraints, all maximum entropy distributions reduce to the microcanonical uniform distribution P_0 described in Sec. II A). This restriction can have two (related) effects: one on the functional form of the maximum entropy distribution and one on the way the distribution connects to the entropy itself and possibly other quantities. The HT and UJK entropies serve as good examples for both effects, as we now show.

For instance, while the general form for the probability distribution that maximizes the HT entropy $S_{c,d}^{(h,g)}[P]$ in trace form ($h(x) = x$) is the exponential of the so-called Lambert-W function[3] $\mathcal{W}(x)$ [8,25], the only admissible form according to our axiom is the one corresponding to the choice $(c, d) = (1, 1)$. With this parameter choice, the $\mathcal{W}(x)$ function reduces to a linear function, so that the maximum entropy probability reduces to the Boltzmann-Gibbs distribution in Eq. (5) [8], consistently with the fact that the only admissible trace-form HT entropy according to our axiom is Shannon entropy, as we have shown above. To obtain a truly generalized maximum-entropy probability, one should therefore consider non-trace-form entropies.

In particular, considering the Rényi entropy $S_q[P]$, which our axiom selects from both the UJK and the HT families, the GMEP can be formulated as the following well-known generalization of the MEP described in Sec. II A. Given an empirically observed value C^* of a (scalar or vector) function $C(G)$ of the unknown microstate G of a system, the least biased inference about G is provided by the distribution P_q that maximizes $S_q[P]$ under the (soft) constraint

$$\langle C \rangle_q \equiv \frac{\sum_{i=1}^{\Omega} p^q(G_i)\, C(G_i)}{\sum_{i=1}^{\Omega} p^q(G_i)} = C^*, \tag{31}$$

which generalizes the usual Shannonian constraint in Eq. (3) (note that $\langle C \rangle_1 = \langle C \rangle$). The quantity $\langle C \rangle_q$ is sometimes called (normalized) q mean and can be regarded as a mean with respect to the so-called escort (or zooming) probability distribution $\tilde{p}(G_i) = p^q(G_i)/\sum_{j=1}^{\Omega} p^q(G_j)$ [7,46]. This q mean has been introduced to extend important properties and relations from the classical (i.e., Shannonian) statistical mechanics to the nonextensive one, including the Legendre structure of thermodynamics, the H theorem and the Ehrenfest theorem [7]. However, from the point of view of statistical inference, whether $\langle C \rangle_q$ is a proper choice for a constraint is a debated issue in the literature, since this quantity might appear to lack a direct interpretation in relation to the available data. We will provide reassurance towards this concern: the use of $\langle C \rangle_q$ leads to a well-defined combination of the available data and, conveniently, regularizes the inference procedure in cases when $\langle C \rangle$ would be unstable. Indeed, $\langle C \rangle_q$ is always finite as soon as the distribution of C is normalizable, even when the ordinary mean $\langle C \rangle$ diverges and the ordinary inference process becomes inapplicable. Ensuring a finite value is crucial in

[3]The Lambert-W function $\mathcal{W}(x)$, which cannot be written in close form, is the solution to the equation $x = \mathcal{W}(x)e^{\mathcal{W}(x)}$. The real solutions are those that are relevant here.

order to estimate the Lagrange multiplier(s) from repeated observations and is especially important in our setting (described later in more detail) where we want to be able to determine q purely from data, without prior knowledge of its value and therefore without knowing beforehand whether the ordinary mean would diverge. We also emphasize that, while in physical systems the natural choice of the constraint(s) is the energy and/or other conserved quantities, in nonphysical ones this choice depends on the nature of the system itself, as it should always capture the most informative properties to achieve the optimal inference. For instance, in network theory, local constraints acting on each node are commonly adopted [47]. A second concern may have arisen in the careful readers who noticed that the UJK entropies are usually derived from the SJ axioms while constraining the ordinary mean value $\langle C\rangle$, not $\langle C\rangle_q$. However, later in this paper we will show that the application of the ML principle to all parameters of the distribution (including q) leads exactly to the same numerical values of the GMEP probability distribution, irrespective of whether $\langle C\rangle$ or $\langle C\rangle_q$ is used (provided both quantities are finite). Finally and profoundly, the use of the q mean restores a complete consistency between the ML principle and the Lagrangian optimization and yields a direct relationship between maximized likelihood and entropy, while the use of the ordinary mean would fail to do so.

To carry out the constrained maximization of $S_q[P]$, we look for the vanishing derivatives of the q Lagrangian

$$\mathcal{L}_q[P] \equiv S_q[P] - \alpha\left[\sum_{i=1}^{\Omega} p(G_i) - 1\right] - \theta\cdot[\langle C\rangle_q - C^*] \tag{32}$$

with respect to P, α, and θ, and assume $q \neq 1$ from now on. The resulting values are denoted as P_q, α_q, θ_q. In particular, setting $\partial\mathcal{L}_q[P]/\partial P|_{P_q} = 0$ (vanishing first component of the derivative) we get

$$0 = \left.\frac{\partial\mathcal{L}_q[P]}{\partial p(G_i)}\right|_{p_q(G_i)} = \frac{q}{1-q}\frac{p_q^{q-1}(G_i)}{\sum_j p_q^q(G_j)} - \alpha - q\,p_q^{q-1}(G_i)\frac{\theta\cdot(C(G_i)-\langle C\rangle_q)}{\sum_j p_q^q(G_j)} \tag{33}$$

for all i from 1 to Ω, from which it is clear that $p_q(G_i)$ depends on θ, as in the case $q = 1$, and additionally on q. The vanishing derivative of $\mathcal{L}_q[P]$ with respect to α (vanishing second component of the derivative) leads to a condition identical to Eq. (6):

$$\left.\frac{\partial\mathcal{L}_q[P_q]}{\partial\alpha}\right|_{\alpha_q} = 0 \Rightarrow \sum_{i=1}^{\Omega} p_q(G_i,\theta) = 1, \tag{34}$$

which can be used to determine α_q by multiplying both sides of Eq. (33) and then summing over i. We then get

$$\alpha_q = \frac{q}{1-q} \quad (q \neq 1), \tag{35}$$

which is the counterpart of Eq. (7). Substituting α_q in (33) and singling out $p_q(G_i)$ yields

$$p_q(G_i,\theta) = \frac{(1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q))_+^{1/(1-q)}}{\left(\sum_{j=1}^{\Omega} p_q^q(G_j,\theta)\right)^{1/(1-q)}}, \tag{36}$$

where we have used the notation $[x]_+^a \equiv 0$ if $x < 0$, while $[x]_+^a \equiv x^a$ otherwise [7]. Note that the denominator of Eq. (36) equals the Uffink functional $U_q[P_q(\theta)]$ and must also equal the generalized partition function

$$W_q(\theta) \equiv \sum_{i=1}^{\Omega}[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)} \tag{37}$$

since $p_q(G_i,\theta)$ is already normalized via the condition in Eq. (35). In other words,

$$W_q(\theta) = \left[\sum_{i=1}^{\Omega} p_q^q(G_i,\theta)\right]^{1/(1-q)} = U_q[P_q(\theta)]. \tag{38}$$

Finally, the maximum entropy probability equals

$$p_q(G_i,\theta) = \frac{[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)}}{W_q(\theta)}, \tag{39}$$

which has the form of a so-called q-exponential or Tsallis [7] distribution. Note that Eqs. (32) and (39) generalize Eqs. (4) and (5), respectively. Moreover note that, if we formally introduce a pseudostate $\tilde{G}$ such that $C(\tilde{G}) = \langle C\rangle_q$, it follows from Eq. (39) that $p_q(\tilde{G},\theta) = 1/W_q(\theta) = 1/U_q[P_q(\theta)]$. Then, from Eq. (38), one can see that:

$$p_q^{q-1}(\tilde{G},\theta) = \sum_{i=1}^{\Omega} p_q^q(G_i,\theta) = U_q^{1-q}[P(\theta)]. \tag{40}$$

We will discuss the relationship between $C(\tilde{G})$ and $C(G^*)$ later.

Before imposing the vanishing third component of the derivative of the q Lagrangian, let us make some general remarks about q exponentials. When $q \to 1$, $p_q(G_i,\theta) \to Z_1^{-1}(\theta)\exp[-\theta\cdot C(G_i)]$, retrieving the Boltzmann-Gibbs distribution in Eq. (5). When $q \neq 1$, the q exponential has nothing to do with the ordinary exponential and actually has power-law tails proportional to $C(G_i)^{1/(1-q)}$ for large values of $C(G_i)$. The presence of these heavy tails, which are widespread in several real-world complex systems [7,30–42], is one of the reasons why q exponentials have attracted interest, their derivation from the maximization of a suitable entropy appearing convenient and parsimonious [7,8]. In the literature, there is some confusion around the fact that q exponentials derive from the maximization of Tsallis entropy given by Eq. (27). While this is certainly true, it is also true that they derive from any of the UJK entropies in Eq. (17): the distribution maximizing $U_q[P]$ necessarily maximizes $f(U_q[P])$ as well, for any monotonic f (indeed, our derivation above started from Rényi entropy). Therefore, the robust empirical support for the q-exponential distribution that has been highlighted in several analyses of empirical data in, e.g., quantum chemistry [29], high-energy physics [30–34], cosmology [35], finance [36], acoustics [39,40], seismology [37], biology [38], and medicine [41,42] cannot be used as support for a specific

member f of the entropy family $f(U_q[P])$. Indeed, there is no direct empirical evidence, which selects a specific entropy in the family, and the main arguments adopted in the literature towards the use of, e.g., Rényi versus Tsallis entropy remain of theoretical or mathematical nature [7,48], such as invoking consistency with some formal framework. In this respect, our approach here can be regarded as an additional theoretical consistency argument to select entropies within families (e.g., Shannon entropy within the Tsallis family) or restricted entropy families within superfamilies (e.g., Rényi entropy within the UJK and HT families). Specifically, the differences among the members of the UJK entropy family arise when the maximum entropy q exponential is put back into the entropy itself. When this happens, the uninformativeness axiom has the important role of selecting Rényi entropy as the member of the family that solves all the inconsistencies discussed in Sec. II F, as we show later in the paper.

What remains to be done is the determination of the parameter θ via the vanishing of the third component of the derivative of the q Lagrangian. It is useful at this point to introduce the reparameterization

$$\psi(\theta) \equiv \frac{\theta}{1+(1-q)\theta\cdot\langle C\rangle_q}, \tag{41}$$

through which it is possible to (formally) remove $\langle C\rangle_q$ from the expression for $p_q(G_i,\theta)$ and get

$$p_q(G_i,\psi) = \frac{[1-(1-q)\,\psi\cdot C(G_i)]_+^{1/(1-q)}}{Z_q(\psi)}, \tag{42}$$

where, denoting the inverse of $\psi(\theta)$ as $\theta(\psi)$,

$$Z_q(\psi) \equiv \sum_{i=1}^{\Omega}[1-(1-q)\,\psi\cdot C(G_i)]_+^{1/(1-q)} \tag{43}$$

$$= \frac{W_q(\theta(\psi))}{[1+(1-q)\theta(\psi)\cdot\langle C\rangle_q]^{1/(1-q)}} \tag{44}$$

is the reparametrized partition function. Note that $Z_q(\psi)\neq W_q(\theta(\psi))$ unless $q\to 1$, in which case $\psi\to\theta$ and $W_1(\theta)\to Z_1(\theta)$. The optimal value ψ_q is determined by the condition

$$\left.\frac{\partial\mathcal{L}[P_q]}{\partial\psi}\right|_{\psi_q} = 0 \Rightarrow \frac{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q)\,C(G_i)}{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q)} = C^* \tag{45}$$

(vanishing third component of the derivative) corresponding to the intended requirement in Eq. (31) and generalizing Eq. (8) to the case $q\neq 1$. One the value ψ_q is determined via the condition above, it can be inserted into Eq. (42) to obtain the final maximum entropy probability distribution $P_q(\psi_q)$.

D. Link with the ML principle and model selection

We now show that the entropy selected by the uninformativeness axiom restores consistency with the ML principle and retains an interpretation for model selection, exactly as in the Shannon case. Both properties are not guaranteed for other entropies. At the same time, we show how to account for multiple independent observations about the same system.

In analogy with Sec. II B, we start with the case $M=1$ and define the ML estimation procedure for the parameter ψ_q as follows:

$$\psi_q^* = \underset{\psi}{\operatorname{argmax}}\,\ell_q(\psi), \qquad \ell_q(\psi)\equiv \ln p_q(G^*,\psi). \tag{46}$$

Requiring $\partial\ell_q(\psi)/\partial\psi|_{\psi_q^*}=0$, one gets

$$\sum_{i=1}^{\Omega} C(G_i)\,p_q^q(G_i,\psi_q^*) = C(G^*)\,p_q^{q-1}(G^*,\psi_q^*) \tag{47}$$

and, dividing both terms by $\sum_G p_q^q(G^*,\psi_q^*)$,

$$\langle C\rangle_q = \frac{C(G^*)\,p_q^{q-1}(G^*,\psi_q^*)}{\sum_G p_q^q(G^*,\psi_q^*)}. \tag{48}$$

One might think that the right-hand side of the above equation is different from the desired value $C^*=C(G^*)$, however this is not the case. Indeed, considering again a pseudostate $\tilde{G}$ such that $C(\tilde{G})=\langle C\rangle_q$ and using Eq. (40), we can rewrite Eq. (48) as

$$\frac{C(\tilde{G})}{C(G^*)} = \frac{1-(1-q)\,\psi_q^*\cdot C(\tilde{G})}{1-(1-q)\,\psi_q^*\cdot C(G^*)}, \tag{49}$$

which leads to $C(\tilde{G})=C(G^*)$. In other words, the value ψ_q^* defined by Eq. (46) coincides with the value ψ_q defined by Eq. (45), i.e., $\psi_q^*\equiv\psi_q$, i.e.,

$$\left.\frac{\partial\ell_q(\psi)}{\partial\psi}\right|_{\psi_q^*} = 0 \Rightarrow \frac{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q^*)\,C(G_i)}{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q^*)} = C^* \tag{50}$$

in analogy with Eq. (45). This means that the ML principle can still be seen as equivalent to the part of the Lagrangian optimization relative to ψ. Moreover, the application of the logarithm to both sides of Eq. (40) leads to

$$\ell_q(\psi_q^*) = -S_q[P_q(\psi_q^*)], \tag{51}$$

showing that, for $M=1$, the log-likelihood of the observation coincides with minus the Rényi entropy. This extends Eq. (11) to the case $q\neq 1$, with the following consequence. If we put ourselves in a model selection framework where we interpret each $P_q(\psi)$ for different values of q as a different model for the same data C^* (for $M=1$) and where each model has a maximized likelihood equal to $P_q(\psi_q^*)$, then the optimal model (if unique) corresponds to the one with the highest value of $P_q(\psi_q^*)$. Thanks to Eq. (51), we can equivalently interpret this ranking of different models according to their maximized likelihood as a ranking based on their realized Rényi entropy: the model with highest maximized likelihood is the one with minimum Rényi entropy. Notably, other entropies of the UJK family, including Tsallis entropy, do not manifest this property. Also the relationship in Eq. (12) generalizes as follows:

$$\mathcal{L}_q[P_q(\psi_q^*)] = S_q[P_q(\psi_q^*)] = -\ell_q(\psi_q^*), \tag{52}$$

relating the value of the Lagrangian attained by $P_q(\psi_q^*)$ to the maximized log-likelihood. Therefore, up to this point, it seems that Rényi entropy retains all the desirable properties of Shannon entropy.

We now consider the case of $M>1$ i.i.d. realizations $\{G_m^*\}_{m=1}^M$ of the system, leading to M independent observations $\{C_m^*\}_{m=1}^M$ of the constraint, where $C_m^*\equiv C(G_m^*)$ for all m. We have already seen in Sec. II B that in this case it is the ML

principle, not the MEP, that identifies how to combine the M observed values. Introducing again the average log-likelihood $\overline{\ell}_q(\psi)$, the ML condition for ψ becomes a straightforward generalization of Eq. (13):

$$\psi_q^* = \underset{\psi}{\text{argmax}}\, \overline{\ell}_q(\psi), \quad \overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^{M} \ln p_q(G_m^*, \psi)}{M}. \tag{53}$$

It is not difficult to show that requiring $\partial \overline{\ell}_q(\psi)/\partial\psi|_{\psi_q^*} = 0$ translates into:

$$\sum_{i=1}^{\Omega} C(G_i)\, p_q^q(G_i, \psi_q^*) = \frac{1}{M} \sum_{m=1}^{M} C(G_m^*)\, p_q^{q-1}(G_m^*, \psi_q^*) \tag{54}$$

or equivalently

$$\langle C \rangle_q = \frac{\sum_{m=1}^{M} C(G_m^*)\, p_q^{q-1}(G_m^*, \psi_q^*)}{M \sum_{i=1}^{\Omega} p_q^q(G_i, \psi_q^*)}, \tag{55}$$

which extends the classical ($q = 1$) result in Eq. (14) to the general, non-Shannon case. We therefore learn that the arithmetic average is no longer the optimal way of combining the M available observations in order to determine the parameter ψ. Indeed, dismissing the arithmetic average makes sense if we recall that, *a priori*, we do not even know whether the first moment of the distribution generating the M values $\{C_m^*\}_{m=1}^M$ is finite. Indeed, the q-exponential distributions that are solution to the GMEP exhibit a power-law behavior for $q \neq 1$. As a consequence, in principle all their moments might diverge, depending on the value of q. Assuming that q is not known beforehand and is rather determined by the inference procedure itself (as we assume later on), it would make no sense at all to use the arithmetic average to constrain the q mean in case of multiple observations, since that average might become infinite in the $M \to \infty$ limit when $q > 3/2$, while the q mean is by construction finite whenever the distribution is normalizable. The same problem might in principle apply to any higher moment $\langle C^n \rangle$ with $n > 1$, while any q-generalized moment $\langle C^n \rangle_q$ evaluated with respect to Eq. (42) converges if q is such that the distribution is normalizable (which is a basic requirement for this procedure to be consistent [7]). The ML estimator determined by Eq. (55) identifies the distribution's parameters, irrespective of the converge of any moment.

An important consequence of the fact that $\langle C \rangle_q$ is no longer equal to the arithmetic mean of the M observations is that in general, for $q \neq 1$ and $M > 1$,

$$S_q[P_q(\psi_q^*)] \neq -\overline{\ell}_q(\psi_q^*), \tag{56}$$

thus failing to generalize Eq. (15) to the case $q \neq 1$ and Eq. (51) to the case $M > 1$. Similarly, Eqs. (12) and (52) do not generalize here. Rather, a relationship that is still valid is

$$S_q[P_q(\psi_q^*)] = -\tilde{\ell}_q(\psi_q^*), \tag{57}$$

where $\tilde{\ell}_q(\psi) \equiv \ln p_q(\tilde{G}, \psi)$ is a sort of pseudolikelihood involving the pseudostate $\tilde{G}$ such that $C(\tilde{G}) = \langle C \rangle_q$ introduced above. Unfortunately, $\tilde{\ell}_q(\psi)$ is no longer equal to the actual log-likelihood $\overline{\ell}_q(\psi)$ based on the M observations. Does this mean that, in presence of multiple i.i.d. observations of the same quantity about a system, the correspondence between log-likelihood and entropy is lost? The answer to this question emerges when looking at a seemingly unrelated problem, i.e., the selection of the optimal value of the entropic parameter q, and is provided in Sec. III E.

E. Inference of the entropic parameter

We now come to the last, and in many ways most crucial, benefit implied by the uninformativeness axiom, namely the possibility of consistently identifying the entropic parameter(s) purely from the data, without postulating *a priori* knowledge about the system, such as scaling laws of the type exemplified by Eqs. (20) and (21) [8,25,26].

To this end, starting directly with the general case $M \geqslant 1$, we invoke again the ML principle and, building on its restored consistency with the estimation of the other parameters of the maximum entropy distribution proven in Eq. (50), extend it to the identification of the entropic parameter(s) themselves. This means that we now turn the model selection procedure we discussed after deriving Eq. (51) (where we compared different models $\{P_q(\psi)\}_q$ in terms of their maximized likelihoods $\{P_q(\psi_q^*)\}_q$) into a single ML parameter estimation procedure, applied directly to the two-parameter distribution $P_q(\psi)$. Indeed, the ML principle treats any parameter agnostically, without specific interpretations, and is therefore unaware of the fact that q and the other structural parameters play different roles in an information-theoretic setting. Considering again Reńyi entropy as the only viable entropy from the UJK and HT families, the ML principle applied also to the entropic parameter q is formally stated as follows:

$$(\psi_{q^*}^*, q^*) = \text{argmax}_{(\psi,q)} \overline{\ell}_q(\psi),$$
$$\overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^{M} \ln p_q(G_m^*, \psi)}{M}. \tag{58}$$

This expression immediately tells us that, once the ML principle is extended to the determination of q, the results we have discussed in Sec. III D represent only one side of the coin. Now, requiring jointly

$$\left.\frac{\partial \overline{\ell}_q(\psi)}{\partial \psi}\right|_{(\psi_{q^*}^*, q^*)} = 0, \quad \left.\frac{\partial \overline{\ell}_q(\psi)}{\partial q}\right|_{(\psi_{q^*}^*, q^*)} = 0, \tag{59}$$

we arrive again at Eq. (55) (with q replaced by q^*) plus the additional condition

$$\sum_{i=1}^{\Omega} p_{q^*}(G_i, \psi_{q^*}^*) \ln[1 - (1 - q^*)\, \psi_{q^*}^* \cdot C(G_i)]$$
$$= \frac{1}{M} \sum_{m=1}^{M} p_{q^*}(G_m^*, \psi_{q^*}^*) \ln[1 - (1 - q^*)\, \psi_{q^*}^* \cdot C(G_m)]. \tag{60}$$

Recalling from Eq. (42) that

$$1 - (1 - q^*)\, \psi_{q^*}^* \cdot C(G_i) = [p_{q^*}(G_i, \psi_{q^*}^*) Z_{q^*}(\psi_{q^*}^*)]^{1-q^*} \tag{61}$$

we obtain the condition

$$\sum_{i=1}^{\Omega} p_{q^*}(G_i, \psi_{q^*}^*) \ln p_{q^*}(G_i, \psi_{q^*}^*) = \frac{\sum_{m=1}^{M} \ln p_{q^*}(G_m^*, \psi_{q^*}^*)}{M}. \tag{62}$$

In other words, the additional ML condition determining q^* requires that the maximized log-likelihood equals minus Shannon entropy, i.e.,

$$S_1[P_{q^*}(\psi^*_{q^*})] = -\overline{\ell}_{q^*}(\psi^*_{q^*}), \tag{63}$$

restoring an analogy with Eq. (15) that appeared to be lost and replaced by Eq. (57) when considering $q \neq 1$. Remarkably, we now realize that, when the ML principle is extended to q, the correspondence with Eq. (15) is not replaced, but rather accompanied by Eq. (57). On one hand, Eq. (63) generalizes to the class of q exponentials the relationship in Eq. (15) that is well-known for the exponential distribution. On the other hand, it does not hold for any value of q and independently on the data, but only for the pair of values $(\psi^*_{q^*}, q^*)$ that maximize the likelihood. In particular, the connection between Shannon entropy and log-likelihood at the specific parameter value $(\psi^*_{q^*}, q^*)$ remains a general result, even for $q \neq 1$ and $M > 1$. This might look quite surprising, because, for $q \neq 1$, the log-likelihood is based on the q-exponential distribution that maximizes Rényi, not Shannon, entropy.

Despite the surprise, the above result makes perfect sense because we have assumed M independent observations. Actually, it solves what would otherwise represent an inconsistency: assuming independent observations does justify Shore and Johnson's original restricted interpretation of axiom *SJ3* and should therefore lead (once all parameters are set to their optimal values) to Shannon entropy as the quantifier of the uncertainty of the data. Indeed the inequality in Eq. (56) should be put in relation with our initial discussion of the axiom *SJ3* about system independence. Recall that assuming that the M values $\{C^*_m\}_{m=1}^M$ come from independent observations is equivalent to assuming that there are M identical and independent copies of the same system, each copy being observed exactly once. Under this assumption of independence, the original reasoning by Shore and Johnson becomes appropriate and one should therefore expect that Shannon entropy, rather than Rényi entropy, is the proper entropy describing the combined system of M copies. Therefore the breakdown of the correspondence between the average log-likelihood and Rényi entropy can be regarded as a symptom of the assumed independence of the M observations. When $M = 1$, we can use Eq. (51) and combine it with Eq. (63) to obtain

$$S_{q^*}[P_{q^*}(\psi^*_{q^*})] = S_1[P_{q^*}(\psi^*_{q^*})] \tag{64}$$

showing that in this particular case the maximum entropy probability distribution returns coinciding values of Shannon and Rényi entropy, even if it maximizes the latter but not the former. This result does not hold in general for $M > 1$. An important consequence of the combination of Eqs. (15) and (63), valid also for $M > 1$, is that the Shannon entropy of the distribution that maximizes Rényi entropy is not larger than that of the exponential distribution, if the parameters are set according to ML:

$$S_1[P_{q^*}(\psi^*_{q^*})] = -\overline{\ell}_{q^*}(\psi^*_{q^*}) \leqslant -\overline{\ell}_1(\psi^*_1) = S_1[P_1(\psi^*_1)], \tag{65}$$

meaning that the optimized q^* exponential achieves, in the Shannon sense, a better compression of the data than the ordinary (Boltzmann-Gibbs) exponential. As we show below for a simple example, $S_1[P_{q^*}(\psi^*_{q^*})]$ and $S_1[P_1(\psi^*_1)]$ might even scale differently with the number of observations (or size of the system), making the inequality (65) particularly relevant for data compression purposes.

The remarkable result in Eq. (63) has an important consequence for the estimation of q^*. In particular, in order to determine both q^* and $\psi^*_{q^*}$, one can consider a range of values for q and, for each value in the range, compute ψ^*_q according to Eq. (55). This produces, for each value of q, a log-likelihood $\overline{\ell}_q(\psi^*_q)$ that is only partially maximized, in the sense that the maximization has been carried out only with respect to ψ_q and not yet with respect to q. Then, among all these partially maximized log-likelihoods, one can select the one with the highest value. This will identify the value q^* and the associated value $\psi^*_{q^*}$, which ultimately correspond to the completely maximized log-likelihood $\overline{\ell}_{q^*}(\psi^*_{q^*})$. Only for this parameter choice $(q^*, \psi^*_{q^*})$, the log-likelihood equals minus Shannon entropy. So from the ML condition Shannon entropy emerges spontaneously: while the probability P_q maximizes Rényi entropy and not Shannon entropy, the latter is the correct entropy for model selection to take independence into account. We stress once again that the physical entropy characterizing the real system may be different from both Shannon and Rényi entropies. Nonetheless, the introduction of our axiom is consistent with an information-theoretic model selection criterion, based on the maximization of Rényi entropy to obtain the functional form of the probability distribution and the ML principle to estimate its parameters, including the entropic one. In order to illustrate the performance of the above approach, we now consider two simple numerical examples.

Our first example is a system described by an observable $C(G)$ taking only positive real values, i.e., $C(G) \in [0, +\infty)$. Moreover, we assume that $\Omega_C = 1$ for all C, meaning that for each value $C(G)$ of the observable there is only one state G that realizes it. Thus, the sums over system states simplify into integrals over the observable values: $\sum_{i=1}^{\Omega} \to \int_0^\infty dC$. The probability distribution resulting from the GMEP is then:

$$p_q(G_i, \psi) = (2-q)\,\psi[1-(1-q)\,\psi \cdot C(G_i)]_+^{\frac{1}{1-q}}, \tag{66}$$

where we have used $Z_q(\psi) = 1/(2-q)\psi$. For different values of ψ and q, we have drawn an i.i.d. sample of $M = 10^3$ realizations from the distribution above, with the aim of inferring the true value of those parameters purely from the data so generated. In particular, we have generated samples from an exponential distribution (i.e., $q_{\text{true}} = 1$), a q-exponential distribution with finite first moment $\langle C \rangle$ ($q_{\text{true}} = 1.3$) and a q-exponential distribution with diverging first moment ($q_{\text{true}} = 1.6$). Figure 1 shows, for the three cases, $\overline{\ell}_q(\psi^*_q)$ (blue line) and $-S_1[P_q(\psi^*_q)]$ (orange line) as functions of q. The black dot indicates the intersection between the two curves, which identifies the estimated value q^* where Eq. (63) is realized. Since the left plot corresponds to $q_{\text{true}} = 1$, it is a standard exponential distribution. In such a case, the two curves intersect only for $q = 1$. By contrast, the other two cases correspond to $q_{\text{true}} \neq 1$ and the two curves intersect in two points, namely $q = 1$ and $q = q_{\text{true}}$. In these cases, both intersections are solutions of Eq. (63), but the solution $q \neq 1$ is the one that corresponds to higher log-likelihood (and lower entropy). The true values of the parameters and their inferred ML estimates

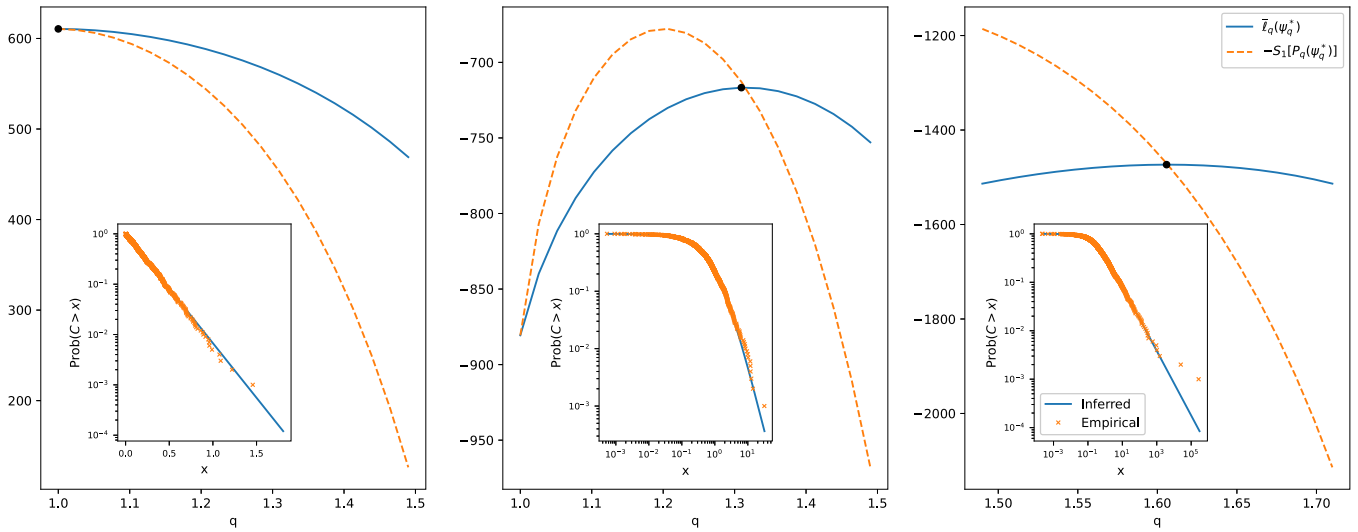

FIG. 1. Comparison between the average partially maximized log-likelihood $\overline{\ell}_q(\psi_q^*)$ (solid line) and minus Shannon entropy $-S_1[P_q(\psi_q^*)]$ (dashed line) as a function of q, for three samples of $M = 10^3$ deviates generated from the probability distribution $P_q(\psi)$ in Eq. (66), and in particular: exponential distribution where $q_{\text{true}} = 1$ and $\psi_{\text{true}} = 5.0$ (left), q-exponential (power-law) distribution with finite first moment where $q_{\text{true}} = 1.3$ and $\psi_{\text{true}} = 3.0$ (center), and q-exponential (power-law) distribution with diverging first moment where $q_{\text{true}} = 1.6$ and $\psi_{\text{true}} = 7.0$ (right). The insets show the comparison between the empirical complementary cumulative distributions of the M realized values (crosses) and the retrieved maximum entropy distribution using the inferred values $(q^*, \psi_{q^*}^*)$ (solid line).

$(q^*, \psi_{q^*}^*)$ are presented in Table I. The estimates $(q^*, \psi_{q^*}^*)$ have been found by solving Eq. (59) for the particular distribution in Eq. (66) (see also Ref. [49] for an equivalent derivation). This example is very simple but explanatory: it shows directly how Shannon entropy plays a role in model selection even when the distribution taken into consideration comes from the GMEP and maximizes Rényi, not Shannon. We also stress once more that, in the last case, constraining the usual mean rather than the q mean would have not been appropriate, since for $q > 1.5$ the usual mean diverges as $M \to \infty$; instead, by using the q average, it becomes possible to consistently characterize the original infinite-mean power-law distribution. Note that, in the real world, the physical entropy characterizing the system producing the data simulated here might be unspecified and, as such, could be different from the information-theoretic one we are using for the inference procedure. Indeed, as stated above, also Tsallis entropy and the HT entropies are naturally associated with power-law distributions. Moreover we would also like to notice that, since the maximized log-likelihood coincides with minus the Shannon entropy, the selected probability distribution could be more compressible than the one obtained setting $q = 1$. An example is shown in the middle panel of Fig. 1: if the probability in Eq. (66) represented the probability distribution of a source generating i.i.d. symbols, then the one selected by our GMEP and ML would be, unsurprisingly, more compressible then the one obtained by setting $q = 1$. Moreover, the precise value $\overline{\ell}_{q^*}(\psi_{q^*}^*)$ coincides with the lower bound of compression, in accordance with Eq. (65). We also note that, if the true process generating the data has finite mean, then the estimation of ψ_1^* is a well-defined problem, otherwise it is not. Figure 2 shows that the Shannon entropy of both distributions is linear in M, with the one referred to the exponential distribution being larger. Moreover, since in this case the data are generated according to a finite-mean distribution, the fluctuations in both $\psi_{q^*}^*$ and ψ_1^* are small. On the other hand, Fig. 3 shows a situation in which the generating process has a diverging mean. While the estimated $\psi_{q^*}^*$ is robust, ψ_1^* has huge fluctuations and results in a superlinear growth of the Shannon entropy. Since $\psi_1^* = M/\sum_{j=1}^M C_j^*$, the large fluctuations arise from the

TABLE I. Comparison of true parameters' values with ML estimates.

q_{true}	ψ_{true}	q^*	$\psi_{q^*}^*$
1.0	5.0	1.0	5.0
1.3	3.0	1.3	2.9
1.6	7.0	1.6	7.3

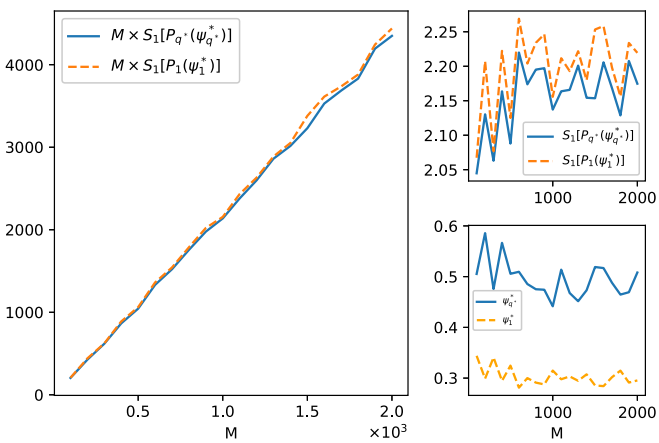

FIG. 2. Data generated according to Eq. (66) with $q_{\text{true}} = 1.2$ and $\psi_{\text{true}} = 0.5$. Left: Shannon entropy. Top right: Shannon entropy per observation. Bottom right: Estimated Lagrange multipliers.

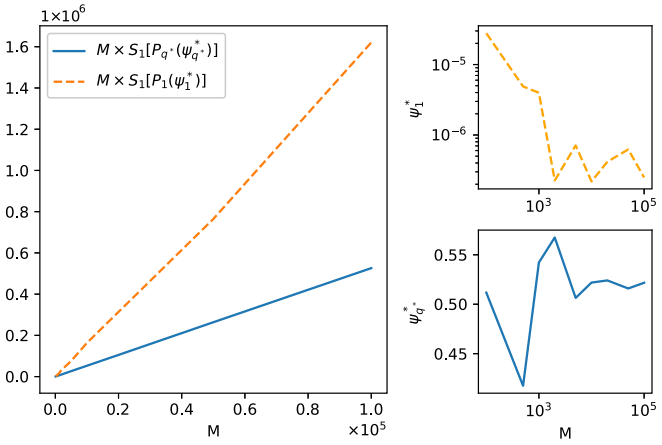

FIG. 3. Data generated according to Eq. (66) with $q_{\text{true}} = 1.7$ and $\psi_{\text{true}} = 0.5$. Left: Shannon entropy. Top right: Estimated ψ_1^*. Bottom right: Estimated $\psi_{q^*}^*$.

fluctuating empirical average of M observations coming from an infinite mean process. Indeed in this particular case, where the observations come from the distribution in Eq. (66) with $q = 1.6$, i.e., $p(C) \sim C^{-\alpha-1}$ with $\alpha = 2/3$, we get that, for large M, the arithmetic average diverges as $M^{-1}\sum_{j=1}^{M} C_j^* \sim M^{1/\alpha-1} = M^{1/2}$. The use of the q average fixes this problem, keeping the estimation of its associated Lagrange multiplier stable. Such divergence is evident if one plugs the estimated probability densities back into the Shannon entropy. In fact, in general, $S_1[P_q(\psi)] = \frac{1}{2-q} - \log((2-q)\psi)$. This leads to a superlinear growth of the entropy if the standard MEP is applied: $S_1[P_1(\psi_1^*)] = 1 - \log\psi_1^* \sim \log M^{1/2}$. Instead, by applying the generalized GMEP, the entropy is $S_1[P_{q^*}(\psi_{q^*}^*)] = \frac{1}{2-q^*} - \log((2-q^*)\psi_{q^*}^*) \sim$ const, since both q^* and $\psi_{q^*}^*$ do not change with M. Incidentally, this example shows that Shannon entropy, even if additive for independent events, can grow nonlinearly in the number of independent observations in case of an anomalous scaling of the relevant Lagrange multiplier(s).

Our second and last example is the simple case of a system characterized by a Bernoulli random variable $C(G)$ taking value $C(G) = 1$ with true underlying probability p_{true}, and value $C(G) = 0$ with probability $1 - p_{\text{true}}$. Constraining the q average yields

$$p_q(G_i, \psi) = \frac{[1 - (1-q)\,\psi \cdot C(G_i)]_+^{1/(1-q)}}{1 + [1 - (1-q)\,\psi]^{1/(1-q)}}. \tag{67}$$

Let us now call $p_q(\psi)$ the probability $p_q(G, \psi)$ when $C(G) = 1$ and $1 - p_q(\psi)$ the probability $p_q(G, \psi)$ when $C(G) = 0$. It is easily verified that

$$\langle C \rangle = p_q(\psi) \tag{68}$$

and

$$\langle C \rangle_q = \frac{p_q^q(\psi)}{p_q^q(\psi) + [1 - p_q(\psi)]^q}. \tag{69}$$

If we now consider M i.i.d. realizations $\{C_m^*\}_{m=1}^M$ of C and apply Eq. (54), we get $p_{q^*}^{q^*}(\psi_{q^*}^*) = p_{q^*}^{q^*-1}(\psi_{q^*}^*) f_1$ where $f_1 = \sum_{m=1}^{M} C_m^*/M$ is the empirical frequency of the observed instances where $C_m^* = 1$. This relation trivially reduces to

$$f_1 = p_{q^*}(\psi_{q^*}^*). \tag{70}$$

Since there are infinite couples of $(\psi_{q^*}^*, q^*)$ that satisfy the ML condition and produce exactly the same maximized log-likelihood, none of them has to be preferred over the other. According to our approach, one finds a result, which recalls the Shannonian case: for a Bernoulli random variable, the parameters of the maximum entropy distribution have to be set so that the estimated probability matches the empirical frequency. This can be done for any value of q and is therefore a degenerate case where no specific value of q can be learned from the data, because the resulting maximum entropy distributions are all identical to each other. This is not unexpected: in fact, what we have done here in practice is trying to capture the properties of a one-parameter binary random variable with a distribution that depends on two parameters. Therefore this example illustrates a situation where the data are not informative enough to infer the entropic parameter.

F. Relation with ordinary average constraints

The SJ axioms are explicitly formulated for constraints that are linear in probability: they not only find a particular functional form for the entropy, but define the whole maximum entropy procedure including the estimation of the Lagrange multipliers. Therefore, one may question whether it makes sense to apply our GMEP procedure, which uses q means, to the UJK family, which derives from the SJ axioms.

In Ref. [21], Jizba and Korbel evaluated explicitly the functional form of the probability distribution resulting from the maximization of $U_{q'}[P]$ with linear constraints. Here, we report their result using our notation. If one constrains the ordinary mean $\langle C \rangle$ and follows the same maximization procedure for $S_{q'}[P]$ as described in Sec. III C, a different maximum entropy distribution $\hat{P}_q(\theta)$ is obtained:

$$\hat{p}_{q'}(G_i) = \frac{[1 - (q'-1)\,\hat{\theta} \cdot (C(G_i) - \langle C \rangle)]_+^{\frac{1}{q'-1}}}{\hat{W}_{q'}(\hat{\theta})}. \tag{71}$$

Following the reparameterization previously introduced, it is also possible to write:

$$\hat{p}_{q'}(G_i, \hat{\psi}) = \frac{[1 - (q'-1)\,\hat{\psi} \cdot C(G_i)]_+^{1/(q'-1)}}{\hat{Z}_{q'}(\hat{\psi})}, \tag{72}$$

where

$$\hat{\psi}(\hat{\theta}) \equiv \frac{\hat{\theta}}{1 + (q'-1)\,\hat{\theta} \cdot \langle C \rangle}. \tag{73}$$

Note that the transformation $q' \to 2 - q$ formally links the two types of constraint. In particular, one can see that

$$\hat{p}_{q'}(G_i, \hat{\psi}) = p_{2-q'}(G_i, \hat{\psi}) = p_q(G_i, \hat{\psi}). \tag{74}$$

The difference between our and UJK approaches lies in the estimation of the Lagrange parameters, and in the fact that UJK assume that q is deducible from the physical properties of the system. They consider a set of observations $\{C_m^*\}_{m=1}^M$ coming from systems that are not assumed to be mutually

independent, and select the Lagrange multiplier in order to satisfy:

$$\langle C\rangle(\hat{\theta}^*) = C^*_{\rm JK}, \tag{75}$$

where $C^*_{\rm JK} = \frac{1}{M}\sum_{m=1}^{M} C^*_m$ is simply the sample (arithmetic) average of the observations. So, they are using the method of moments as estimation technique. It follows that their probability distribution becomes:

$$\hat{p}_{q'}(G_i) = \frac{[1-(q'-1)\,\hat{\theta}^*\cdot(C(G_i)-C^*_{\rm JK})]_+^{\frac{1}{q'-1}}}{\hat{W}_{q'}(\hat{\theta}^*)}. \tag{76}$$

With our approach we consider instead a different situation. We imagine that our set of observations comes from replicas of the same system. Thus, we are considering an ensemble of systems which we know are independent of each other, but could internally have correlations which are better captured by generalized entropies (and, consequentially, nonfactorizable probabilities). In other words, we are addressing a hierarchical property of entropy, which arises from the scale dependence of correlations. Even if our GMEP procedure does not satisfy *SJ4* (subset independence) axiom due to the presence of nonlinear constraints, it is consistent with the fact that the resulting probabilities should behave differently depending on the scale of the conditioning subset in *SJ4*. Moreover, it is possible to establish a relationship between the two approaches, showing that our GMEP-ML procedure returns a probability distribution which is equivalent to the one UJK would obtain by considering a deformed $C^*_{\rm JK}$. The probability distribution resulting from our approach has the form:

$$p_q(G_i,\psi^*) \propto [1-(1-q)\,\psi^*\cdot C(G_i)]_+^{1/(1-q)}, \tag{77}$$

where ψ^* is related to θ^* through Eq. (41) and is here a numerical value satisfying Eq. (54). It is possible to rewrite ψ^* as a function of the linear average of C with respect to P_q:

$$\psi^* = \frac{\hat{\theta}^*}{1+(1-q)\hat{\theta}^*\cdot\langle C\rangle}. \tag{78}$$

Plugging ψ^* back into Eq. (77), we get:

$$p_q(G_i,\hat{\theta}^*(\psi^*)) \propto [1-(1-q)\,\hat{\theta}^*\cdot(C(G_i)-\langle C\rangle)]_+^{1/(1-q)}, \tag{79}$$

which is equivalent to Eq. (76) under the transformation $q \to 2-q'$ and by setting $C^*_{\rm JK} = \langle C\rangle \neq M^{-1}\sum_{m=1}^{M} C^*_m$. The important difference with respect to the JK approach is that $\langle C\rangle$ is not the sample average of the observation, but the theoretical mean value calculated with respect to the P_q resulting from our approach, based on GMEP and ML estimation. This is neither surprising nor in contradiction with the SJ axioms or the JK approach: while the sample average is a good estimator when one wants to be completely unbiased with respect to the possible correlations among different systems, it fails when one knows that at certain scales probability distributions eventually factorize (independent systems), but still wants to allow for dependencies at finer scales (dependent subsystems within each system).

In order to complete the discussion, in the following we describe how our approach behaves when the linear average is constrained. According to the GMEP, we maximize the Lagrangian:

$$\mathcal{L}'_{q'}[\hat{P}] = S_{q'}[\hat{P}] - \alpha\left[\sum_{i=1}^{\Omega}\hat{p}(G_i)-1\right] - \hat{\theta}\cdot[\langle C\rangle - C^*]. \tag{80}$$

We obtain then the probability distribution in Eq. (71), which is equivalent to Eq. (72), and use the ML principle to estimate $\hat{\psi}^*$ and q'^*. It is easy to show that ML on $\hat{\psi}$ leads to a condition, which, under the $q' \to 2-q$ transformation, is equivalent to Eq. (54):

$$\sum_{i=1}^{\Omega} C(G_i)\,\hat{p}_{q'}^{2-q'}(G_i,\hat{\psi}^*_{q'}) = \frac{1}{M}\sum_{m=1}^{M} C(G^*_m)\,\hat{p}_{q'}^{1-q'}(G^*_m,\hat{\psi}^*_{q'}). \tag{81}$$

The same holds for the estimation of the entropic parameter, which satisfy Eq. (63):

$$S_1[\hat{P}_{q'^*}(\hat{\psi}^*_{q'^*})] = -\overline{\ell}_{q'^*}(\hat{\psi}^*_{q'^*}). \tag{82}$$

Imagine now to have a set of observations and to estimate the parameters of $P_q(\psi)$ and $\hat{P}_{q'}(\hat{\psi})$, obtained respectively constraining $\langle C\rangle_q$ and $\langle C\rangle$. We would obviously get that if $q = q^*$, then $q'^* = 2-q^*$ and that $\psi^* = \hat{\psi}^*$. So, according to Eq. (74), we would obtain the same probability distribution in both cases. However, the consistency between the GMEP and the ML is maintained only when the q-average is constrained. In fact, considering a single observation C^* corresponding to the state G^*, in the first case we have:

$$\mathcal{L}_{q^*}[P_{q^*}(\theta^*)] = S_{q^*}[P_{q^*}(\theta^*)] = -\ln p_{q^*}(G^*,\theta^*).$$

In the second case, instead:

$$\begin{aligned}\mathcal{L}'_{2-q^*}[P_{q^*}(\hat{\theta}^*)] &= S_{2-q^*}[P_{q^*}(\hat{\theta}^*)] - \hat{\theta}^*\cdot[\langle C\rangle - C^*]\\ &\neq -\ln p_{q^*}(G^*,\hat{\theta}^*)\\ &\neq -\ln p_{2-q^*}(G^*,\hat{\theta}^*).\end{aligned}$$

Thus, the correspondence between (minus) the log-likelihood and the Lagrangian is valid only in the first case, as well the one with the considered Rényi entropy (i.e., the one we maximize in the GMEP).

IV. CONCLUSIONS

A large body of literature has discussed the generalized axiomatic definition of entropy deriving from the relaxation (or unrestricted interpretation) of some of the SK and SJ axioms (in particular, *SK4* and *SJ3*). It is known that, when generalized in that way, the definition of entropy leads to parametric entropy families where a specific value of the entropic parameter(s) usually retrieves the ordinary Shannon functional. In a maximum entropy approach, each entropy family leads to a corresponding family of maximum entropy probability distributions, indexed again by the entropic parameter(s), that provide the least biased inference about a system for which only limited information is available, in the form of empirical observations of a quantity treated as a soft constraint. Unfortunately, when the estimated maximum entropy distribution is put back into its defining generalized entropy, a number of inconsistencies typically arise, including incompatibility with the ML principle, impossibility of determining the value of

the entropic parameter(s) purely from empirical data, and disconnection from Shannon entropy when multiple independent observations of the same system are available.

In this paper, based on the fact that every member of an entropy family is ultimately intended as a quantification of the uncertainty encoded in the input probability distribution, we have introduced an uninformativeness axiom demanding that the maximally uncertain (i.e., uniform) probability distribution should always return the same (maximal) value of the entropy, irrespective of the value of the entropic parameters. This simple axiom implies that all entropies take values within the same interval $[0, \ln \Omega]$, where Ω is the number of possible (unconstrained) microstates of the system, thereby equipping generalized entropies with a universal scale and meaning. The axiom considerably restricts the admissible members of entropy families. In particular, for both the UJK and HT entropies, the axiom selects only Rényi entropy as viable. A notable counterexample, dismissed by the axiom, is Tsallis entropy. From an inferential point of view, the axiom guarantees that completely uninformative data (or equivalently the complete absence of empirical information) cannot be used to learn the value of entropic parameters. At the same time we have showed that, when informative data are available, a straightforward extension of the ML principle leads to the optimal estimation of the entropic parameter(s), purely from empirical observations and without making any assumptions.

The resulting generalized ML approach couples the determination of the entropic parameters with that of the other structural parameters (Lagrange multipliers) of the maximum entropy distribution. In particular, while the ML condition for the Lagrange multipliers indicates which specific combination of M independent observations should be put equal to the generalized mean value of the constraint, the one for the entropic parameters coincides with the equality between maximized log-likelihood and minus Shannon entropy. This remarkable result shows that the connection between Shannon entropy and log-likelihood holds true also for generalized entropies (for the appropriate ML value of the entropic parameter) and is consistent with the assumed independence of the M observations. When $M = 1$, the maximum entropy probability returns coinciding values of Rényi and Shannon entropies, even if it maximizes the former but not the latter. For multiple independent observations ($M > 1$), the connection between log-likelihood and Shannon entropy remains, while the connection with Rényi entropy disappears, as a result of independence. Therefore the log-likelihood, when maximized also over the entropic parameters, automatically finds, in the parametric family of entropies constructed under the GMEP, the optimal one to be used for model fitting and selection. This makes the GMEP even sharper in identifying the unbiased maximum-entropy probability distribution. We believe that the introduction of the uninformative axiom has beneficial effects for statistical inference and its many applications, offering a way of constructing generalized entropies that have still controllable and consistent properties.

ACKNOWLEDGMENTS

This work is supported by the Dutch Econophysics Foundation (Stichting Econophysics Leiden, the Netherlands) and by the European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR), projects "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics", Grant No. IR0000013 (No. 3264, 28/12/2021), and "Reconstruction, Resilience and Recovery of Socio-Economic Networks" RECON-NET EP_FAIR_005 - PE0000013 "FAIR" - PNRR M4C2 Investment 1.3. This work is also supported by the European Union under the scheme HORIZON-INFRA-2021-DEV-02-01–Preparatory phase of new ESFRI research infrastructure projects, Grant Agreement No.101079043, "SoBigData RI PPP: SoBigData RI Preparatory Phase Project".

[1] R. Clausius, On the application of the mechanical theory of heat to the steam-engine, Philos. Mag. J. Sci. **12**, 241 (1856).

[2] L. Boltzmann, *Über Die Beziehung Zwischen Dem Zweiten Hauptsatze Des Mechanischen WäRmetheorie Und Der Wahrscheinlichkeitsrechnung, Respective Den Sätzen Über Das Wärmegleichgewicht* (Kk Hof-und Staatsdruckerei, 1877)

[3] J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (Courier Corporation, New York, 2014).

[4] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).

[5] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).

[6] K. P. Murphy, *Probabilistic Machine Learning: An Introduction* (MIT Press, Cambridge, 2022).

[7] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World* (Springer, Berlin, 2009).

[8] S. Thurner, R. Hanel, and P. Klimek, *Introduction to the Theory of Complex Systems* (Oxford University Press, Oxford, 2018).

[9] J. M. Amigó, S. G. Balogh, and S. Hernández, A brief review of generalized entropies, Entropy **20**, 813 (2018).

[10] A. M. Lopes and J. A. T. Machado, A review of fractional order entropies, Entropy **22**, 1374 (2020).

[11] J. Korbel, R. Hanel, and S. Thurner, Classification of complex systems by their sample-space scaling exponents, New J. Phys. **20**, 093007 (2018).

[12] S. Thurner, B. Corominas-Murtra, and R. Hanel, Three faces of entropy for complex systems: Information, thermodynamics, and the maximum entropy principle, Phys. Rev. E **96**, 032124 (2017).

[13] J. Karmeshu, *Entropy Measures, Maximum Entropy Principle and Emerging Applications* (Springer, Berlin, 2003), Vol. 119.

[14] T. Squartini and D. Garlaschelli, *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics* (Springer, Berlin, 2017).

[15] D. Garlaschelli and M. I. Loffredo, Maximum likelihood: extracting unbiased information from complex networks, Phys. Rev. E **78**, 015101(R) (2008).

[16] D. R. A. Kenneth P. Burnham, *Model Selection and Multimodel Inference* (Springer, New York, 2002).

[17] P. D. Grünwald, I. J. Myung, and M. A. Pitt, *Advances in Minimum Description Length: Theory and Applications* (MIT Press, Cambridge, 2005).

[18] A. Khinchin, *Mathematical Foundations of Information Theory* (Dover, New York, 1957).

[19] J. Shore and R. Johnson, Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, IEEE Trans. Inf. Theory **26**, 26 (1980).

[20] J. Uffink, Can the maximum entropy principle be explained as a consistency requirement? Studies Hist. Philos. Sci. B **26**, 223 (1995).

[21] P. Jizba and J. Korbel, Maximum entropy principle in statistical inference: case for non-shannonian entropies, Phys. Rev. Lett. **122**, 120601 (2019).

[22] P. Jizba and J. Korbel, When shannon and khinchin meet shore and johnson: Equivalence of information theory and statistical inference axiomatics, Phys. Rev. E **101**, 042126 (2020).

[23] A. N. Kolmogorov and G. Castelnuovo, *Sur La Notion De La Moyenne* (G. Bardi, tip. della R. Accad. dei Lincei, Rome, 1930).

[24] M. Nagumo, Über eine klasse der mittelwerte, in *Japanese Journal of Mathematics: Transactions and Abstracts* (The Mathematical Society of Japan, Tokyo, 1930), Vol. 7, pp. 71–79.

[25] R. Hanel and S. Thurner, A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions, Europhys. Lett. **93**, 20006 (2011).

[26] S. G. Balogh, G. Palla, P. Pollner, and D. Czégel, Generalized entropies, density of states, and non-extensivity, Sci. Rep. **10**, 15516 (2020).

[27] A. Plastino, H. Miller, and A. Plastino, General thermostatistical formalisms based on parameterized entropic measures, Continuum Mech. Thermodyn. **16**, 269 (2004).

[28] A. G. Bashkirov, Maximum rényi entropy principle for systems with power-law hamiltonians, Phys. Rev. Lett. **93**, 130601 (2004).

[29] R. Wild, M. Nötzold, M. Simpson, T. D. Tran, and R. Wester, Tunnelling measured in a very slow ionmolecule reaction, Nature **615**, 425 (2023).

[30] C.-Y. Wong, G. Wilk, L. J. Cirto, and C. Tsallis, *Possible Implication of a Single Nonextensive pt Distribution for Hadron Production in High-Energy pp Collisions*, in EPJ Web Conf. (EDP Sciences, 2015), Vol. 90, p. 04002.

[31] C.-Y. Wong, G. Wilk, L. J. L. Cirto, and C. Tsallis, From qcd-based hard-scattering to nonextensive statistical mechanical descriptions of transverse momentum spectra in high-energy p p and p p collisions, Phys. Rev. D **91**, 114027 (2015).

[32] D. B. Walton and J. Rafelski, Equilibrium distribution of heavy quarks in fokker-planck dynamics, Phys. Rev. Lett. **84**, 31 (2000).

[33] A. Deppman, E. Megías, and D. P. Menezes, Fractals, nonextensive statistics, and QCD, Phys. Rev. D **101**, 034019 (2020).

[34] E. Megias, A. Deppman, R. Pasechnik, and C. Tsallis, Comparative study of the heavy-quark dynamics with the fokker-planck equation and the plastino-plastino equation, Phys. Lett. B **845**, 138136 (2023).

[35] P. Jizba and G. Lambiase, Tsallis cosmology and its applications in dark matter physics with focus on icecube high-energy neutrino data, Europhys. J. C **82**, 1123 (2022).

[36] J. Ludescher and A. Bunde, Universal behavior of the interoccurrence times between losses in financial markets: Independence of the time resolution, Phys. Rev. E **90**, 062809 (2014).

[37] C. G. Antonopoulos, G. Michas, F. Vallianatos, and T. Bountis, Evidence of q-exponential statistics in greek seismicity, Physica A **409**, 71 (2014).

[38] M. I. Bogachev, A. R. Kayumov, and A. Bunde, Universal internucleotide statistics in full genomes: A footprint of the dna structure and packaging? PLoS ONE **9**, e112534 (2014).

[39] A. Greco, C. Tsallis, A. Rapisarda, A. Pluchino, G. Fichera, and L. Contrafatto, Acoustic emissions in compression of building materials: Q-statistics enables the anticipation of the breakdown point, Eur. Phys. J.: Spec. Top. **229**, 841 (2020).

[40] S. C. Vinciguerra, A. Greco, A. Pluchino, A. Rapisarda, and C. Tsallis, Acoustic emissions in rock deformation and failure: New insights from q-statistical analysis, Entropy **25**, 701 (2023).

[41] D. M. Abramov, C. Tsallis, and H. S. Lima, Neural complexity through a nonextensive statistical–mechanical approach of human electroencephalograms, Sci. Rep. **13**, 10318 (2023).

[42] R. J. Al-Azawi, N. M. Al-Saidi, H. A. Jalab, H. Kahtan, and R. W. Ibrahim, Efficient classification of covid-19 ct scans by using q-transform model for feature extraction, PeerJ Computer Science **7**, e553 (2021).

[43] A. Rényi, On measures of entropy and information, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics* (University of California Press, Berkeley, 1961), Vol. 1, pp. 547–561.

[44] C. Tsallis, Possible generalization of Boltzmann-Gibbs Statistics, J. Stat. Phys. **52**, 479 (1988).

[45] Q. Zhang and D. Garlaschelli, Strong ensemble nonequivalence in systems with local constraints, New J. Phys. **24**, 043011 (2022).

[46] C. Beck and F. Schögl, *Thermodynamics of Chaotic Systems: An Introduction*, Vol. 4 (Cambridge University Press, Cambridge, 1995).

[47] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli, The statistical physics of real-world networks, Nature Rev. Phys. **1**, 58 (2019).

[48] S. Umarov and T. Constantino, *Mathematical Foundations of Nonextensive Statistical Mechanics* (World Scientific, Singapore, 2022).

[49] C. R. Shalizi, Maximum likelihood estimation for q-exponential (Tsallis) distributions, arXiv:math/0701854.