Nonadditive Entropies Yield Probability Distributions with Biases not Warranted by the Data

Steve Pressé,[1,*] Kingshuk Ghosh,[2] Julian Lee,[3] and Ken A. Dill[4]

[1]*Indiana University—Purdue University Indianapolis, Indianapolis, Indiana 46202, USA*
[2]*Department of Physics and Astronomy, University of Denver, Denver, Colorado 80208, USA*
[3]*Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea*
[4]*Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, USA*

(Received 25 June 2013; published 1 November 2013)

Different quantities that go by the name of entropy are used in variational principles to infer probability distributions from limited data. Shore and Johnson showed that maximizing the Boltzmann-Gibbs form of the entropy ensures that probability distributions inferred satisfy the multiplication rule of probability for independent events in the absence of data coupling such events. Other types of entropies that violate the Shore and Johnson axioms, including nonadditive entropies such as the Tsallis entropy, violate this basic consistency requirement. Here we use the axiomatic framework of Shore and Johnson to show how such nonadditive entropy functions generate biases in probability distributions that are not warranted by the underlying data.

DOI: 10.1103/PhysRevLett.111.180604 PACS numbers: 05.20.−y, 02.50.Tt, 89.70.Cf

A problem of broad interest across the sciences is to infer the mathematical form of a probability distribution given limited data [1]. For instance, we may be given limited information on an equilibrium system—say, its average energy—from which we must predict the mathematical form of the full energy probability distribution. In this classic example, the distribution used in statistical mechanics is the exponential Boltzmann distribution.

In many cases—including the case above—limited data are consistent with many possible models for a probability distribution. How should we select the ''best'' model that fits the data? By model, we mean a set of probablities $\{p_k\}$ for the outcomes k of an experiment. One way to select the model is to eliminate candidate models by supplementing the data with additional assumptions [2–5,5,6]. A common additional assumption is based on choosing the model that has the largest entropy. This is the basis for the variational principle called maximum entropy (MaxEnt) [7].

The MaxEnt approach has its historical roots in the work of Boltzmann in equilibrium statistical physics. Later, Shannon and Jaynes showed that picking the model with the largest entropy was analogous to maximizing the uncertainty H of the model [7–10]. In particular, Jaynes drew the connection between statistical mechanics and Shannon's work on information theory by showing that, since Shannon's $H = -\sum p_k \log p_k$ coincides with the Boltzmann-Gibbs (BG) entropy, statistical mechanics could be treated as an inference problem [9,10]. Shannon and Jaynes and others justified the specific mathematical form $H = -\sum p_k \log p_k$ on the basis of abstract properties of H itself, such as satisfying a composition property [7]. Others justified the form for H using the Khinchin axioms [11,12] or the Fadeev postulates [13], for example.

In contrast to these methods for deriving H on the basis of H's properties, Shore and Johnson (SJ) [14] showed that MaxEnt was the only consistent recipe for drawing self-consistent inferences from data. SJ only asserted, by means of four axioms, that any variational procedure must yield a unique probability distribution that satisfies the rules of addition and multiplication for independent probabilities if data do not couple the probabilities for different events. The arguments of SJ are quite general as they do not assign explicit meaning—and, in particular, any thermodynamic meaning—to H itself. Thus, H can be used as a variational function to discriminate between models across a broad range of problems.

However, in recent years other mathematical functions of $\{p_k\}$ also called entropies [12,13,15–21]—and, more broadly, regularization schemes [22] of which entropy maximization is a special type—have been used to infer complex models, often power laws, from data. In general, these entropies violate one or more of SJ's axioms and, as a result, may be nonadditive. In contrast, the BG entropy is additive in the sense that, for two independent systems A and B, the value of the BG entropy H for the combined system satisfies $H(A, B) = H(A) + H(B)$.

Nonadditive entropies have been of particular interest because they are commonly invoked when microscopic components of a system have long-range interactions. These unconventional entropies satisfy different properties from those of BG. As an example, according to Tsallis, Gell-Mann, and Sato [23], the Tsallis entropy [12,16,18,23,25] of a scale-invariant system is extensive. While nonadditive entropies do not satisfy the additivity rule, they may instead satisfy a ''pseudoadditivity rule'' [24], $H(A, B) = H(A) + H(B) + \epsilon H(A)H(B)$, where ϵ is a measure of the deviation from additivity. Nonadditive forms of the entropy function used within variational principles can preferentially select models having power law distributions [18,25] which arise from a variety of natural and social systems.

Nonadditive entropies have been criticized on the basis, for example, that the Tsallis parameter q [26,27] (related to the ϵ presented earlier) is often chosen by fitting data, rather than by some first principle [25]. Furthermore, unconventional averages must often be used to constrain nonadditive entropies [28–33] to assure the convexity of those functions if they are to be used to infer a unique model.

In regard to these criticisms, if the matter at hand concerns situations in which a full distribution of data is already known—and thus q could be fit to data—then it is fair to ask whether there is a need for any variational principle for selecting a model in the first place. This raises the question of how to justify—and when to use—BG versus nonadditive entropies.

The question of interest here is how nonadditive entropies can be justified within SJ's axiomatic framework. This framework is not about specifying properties of H based on physical (and often thermodynamic) properties of an entropy, but rather about making self-consistent inferences from data without imposing structure on a model which is not warranted by the data itself. Thus SJ's axioms are stronger conditions than additivity conditions on H.

We use SJ's reasoning to shed light on what variant of the basic logical consistency requirements is necessary to derive an alternate formula for the entropy. We first review SJ's axioms and show how BG's H follows from the product rule $p_{ij} = u_i v_j$ in the absence of data coupling events i and j. We will then show what rules of probability would be required instead in order to justify the Tsallis entropy as well as other entropies. In particular, we will show from SJ that the Tsallis entropy can only be justified if events i and j were to have the following joint probability, $p_{ij}^{q-1} = (u_i^{q-1} + v_j^{q-1} - 1)$—presupposed in the absence of data coupling events i and j—rather than $p_{ij} = u_i v_j$.

Shore and Johnson axioms.—SJ considered the problem of extracting a model from data using a variational function $H(\{p_k\})$. The model $\{p_k^*\}$ is the one that gives the maximum of

$$H(\{p_k\}) - \lambda\left(\sum_k p_k a_k - \bar{a}\right) \tag{1}$$

with respect to $\{p_k\}$ and the Lagrange multiplier λ. See Refs. [30–33] for a discussion of constraints. Here the data are imposed as a constraint on the quantity a, where $\bar{a}$ is the measured average. For simplicity, we considered here only a single equality constraint.

SJ gave four axioms that must be satisfied by the maximum of the function given by Eq. (1) on the basis of requiring that any inference drawn from this function be self-consistent. These four axioms determine the form of H.

(1) Uniqueness says that the function $H(\{p_k\})$ must be convex, so that there will only be a single maximum, i.e., a single set of values $\{p_k^*\}$.

(2) Coordinate system invariance says that predictions made from an inference should be independent of the choice of coordinate system. It is relevant when the probabilities are continuous functions and determining the dependence of H on the prior over p_k.

(3) Subset independence says that if probability p_k of bin k increases by δp and the probability p_j of bin j correspondingly decreases by δp, then no other bins are affected by the change. Subset independence yields the relationship

$$H = \sum_k f(p_k) + C, \tag{2}$$

where C is a constant independent of p_k.

(4) System independence says that bringing together two systems having probabilities $\mathbf{u} = \{u_i\}$ and $\mathbf{v} = \{v_j\}$ gives new bins that have probability $\mathbf{p} = \mathbf{u} \times \mathbf{v}$, where $p_{ij} = u_i v_j$. The systems are considered independent if constraints on the data do not couple them. Consider a combined system with two decoupled constraints, one on u_i (which is $\sum_{i,j} p_{ij} a_i - \bar{a} = \sum_i u_i a_i - \bar{a} = 0$) and another on v_j (which is $\sum_{i,j} p_{ij} - \bar{b} = \sum_j v_j b_j - \bar{b} = 0$).

The maximum of

$$H(\mathbf{p}) - \lambda_a\left(\sum_{i,j} p_{ij} a_i - \bar{a}\right) - \lambda_b\left(\sum_{i,j} p_{ij} b_j - \bar{b}\right) \tag{3}$$

with respect to $p_{ij} = u_i v_j$ satisfies

$$f'(p_{ij}) - \lambda_a a_i - \lambda_b b_j = 0. \tag{4}$$

Taking two derivatives of the above (one with respect to u_i and another with respect to v_j) yields

$$f''(p_{ij}) + p_{ij} f'''(p_{ij}) = 0. \tag{5}$$

We define $f''(p_\alpha) \equiv g(p_\alpha)$, where $\alpha \equiv (i, j)$. Then Eq. (5) reduces to $g(p_\alpha) + p_\alpha g'(p_\alpha) = 0$. The solution is $g(p_\alpha) = -1/p_\alpha$. It follows that $f(p_\alpha) = -p_\alpha \log p_\alpha + p_\alpha$ and $H = -\sum_\alpha p_\alpha \log p_\alpha + C$, where all additional constants have been absorbed into C.

The derivation above shows how the BG formula follows from the axioms of SJ. Intuitively, SJ's axioms 3 and 4 take as definitions the fact that events are independent unless the data couple them and the probabilities for independent events satisfy the usual rules of addition and multiplication for such probabilities. This explains why the BG entropy is additive.

However, not all physical systems are additive; often cited as counterexamples are systems having long-ranged interactions [18]. The question is, how should nonadditivity be built into a model? One route has been to redefine entropy and replace it with a form which violates axiom 4, system independence—the law of multiplication of probability for independent events, $p_k = u_i v_j$. Here we demonstrate the logical consequences that follow from redefining the entropy.

The Tsallis entropy is defined as

$$H = \frac{K}{1-q}\left(\sum_k p_k^q - 1\right). \tag{6}$$

This expression satisfies subset independence, axiom 3, but does not satisfy system independence, axiom 4. What functional form for $p_{ij} = p(u_i, v_j)$ yields the Tsallis entropy? To answer this question, we repeat steps analogous to those in Eqs. (3)–(5), except now we treat p_{ij} as a general function of u_i and v_j and $f(p_{ij})$ is given by the form Eq. (6). This gives

$$(2-q)^{-1} p_{ij} \frac{\partial^2 p_{ij}}{\partial u_i \partial v_j} = \frac{\partial p_{ij}}{\partial u_i}\frac{\partial p_{ij}}{\partial v_j}. \tag{7}$$

Equation (7) is a differential equation satisfied by the joint probability in the Tsallis entropy. As a check, we can see that for $q = 1$—when the Tsallis entropy reduces to the BG entropy—the expression is exactly satisfied for $p_{ij} = u_i v_j$, as expected. The constant $(2-q)^{-1}$ in Eq. (7) describes the deviation from independence [often one speaks of deviation from independence in terms of the q additivity of the Tsallis entropy (as opposed to normal additivity of entropy in statistical mechanics)] [19,21,24].

We now solve Eq. (7).

Step 1.—Substitute $p_{ij} = h_{ij}^x$, where x is a number, into Eq. (7). After some algebraic rearrangement, this yields

$$-(2-q)^{-1} x h_{ij} \frac{\partial^2 h_{ij}}{\partial u_i \partial v_j} = [x(x-1)(2-q)^{-1} - x^2]\frac{\partial h_{ij}}{\partial u_i}\frac{\partial h_{ij}}{\partial v_j}. \tag{8}$$

We select x such that $[x(x-1)(2-q)^{-1} - x^2] = 0$. The nontrivial solution to this quadratic equation is $x = 1/(q-1)$ (the trivial solution is $x = 0$). Plugging $x = 1/(q-1)$ into Eq. (8), we have

$$\frac{\partial^2 h_{ij}}{\partial u_i \partial v_j} = 0. \tag{9}$$

Equation (9) is solved by $h_{ij} = \phi_1(u_i) + \phi_2(v_j)$ with ϕ_1 and ϕ_2 yet to be determined. Since both p_{ij} and h_{ij} must be symmetric functions of u_i and v_j, then $\phi_1 = \phi_2 \equiv \phi$. We therefore have

$$p_{ij}^{q-1} = h_{ij} = \phi(u_i) + \phi(v_j). \tag{10}$$

Step 2.—In order to determine the function $\phi(x)$, we rewrite it as

$$\phi(x) = g(x)^{q-1} - 1/2 \tag{11}$$

without loss of generality, so that Eq. (10) takes the form

$$p_{ij} = [g(u_i)^{q-1} + g(v_j)^{q-1} - 1]^{1/(q-1)}. \tag{12}$$

The leading order expansion of (12) in $q-1$ is

$$p_{ij} = g(u_i)g(v_j) - (q-1)g(u_i)g(v_j)\log g(u_i)\log g(v_j) + O[(q-1)^2], \tag{13}$$

and from the condition that the composition rule reduces to the product rule in the limit of $q \to 1$, we get $g(x) = x$. Substituting it back into Eq. (12), we get

$$p_{ij} = (u_i^{q-1} + v_j^{q-1} - 1)^{1/(q-1)}. \tag{14}$$

Eq. (14) can also be written in terms of the q product defined by Tsallis [18]. Furthermore, Eq. (7) gives us the choice to select a multiplicative constant that can be determined by normalization of the joint probability, p_{ij}.

Fundamentally, the spurious correlations between events, which consist of all terms beyond the first term in Eq. (13), emerge because the Tsallis entropy violates SJ's axiom 4. This axiom specifically requires that in the absence of couplings between events i and j, the model inferred using the BG entropy should satisfy the normal rules of multiplication of probability ($p_{ij} = u_i v_j$).

Many entropies—beyond the Tsallis entropy—also violate axiom 4. These entropies therefore generate spurious correlations not warranted by the data even if they are additive in the sense $H(\{p_{ij}\}) = H(\{u_i\}) + H(\{v_j\})$. In other words, axiom 4 is a stronger statement than is the statement that H's add.

For instance, consider the Burg entropy [34] $K\sum_k \log p_k$, which is additive in the sense described above. The Burg entropy satisfies SJ's axiom 3 but violates axiom 4. For this entropy, the system dependence relationship still deviates from the rule of multiplication of probability for independent events:

$$p_{ij}^{-1} = u_i^{-1} + v_j^{-1} - 1. \tag{15}$$

Equations (14) and (15) underscore the profound consequences that result from altering the form of the entropy used in model inference. SJ's axioms assure us that the BG form of the entropy enforces a model distribution which is as featureless as possible. According to SJ's framework, couplings between events—or more broadly, structure in a model—arise in one of two ways. Either the couplings in the data explicitly give rise to correlations between events i and j or the prior over the $\{p_k\}$, the set $\{q_k\}$ which can be thought of as a hyperprior, gives rise to structure beyond what is present in the data. Thus, application of the BG entropy ensures that inferences do not go beyond what is in the data or $\{q_k\}$.

However, nontraditional entropies, which violate SJ's axiom 4, are inconsistent with the probability relationship $p_{ij} = u_i v_j$ even in the absence of any evidence of coupling between events i and j. While entropy priors, such as the Tsallis entropy, can readily infer power law distributions for $\{p_k\}$, they impose structure in a model that goes beyond what is known from the data. Here Eqs. (14) and (15) derive this additional structure imposed by these entropies

on a model explicitly. We conclude by adding that it is possible to infer power laws within a principle of maximizing the BG entropy by constraining just one average: Mandelbrot [35] showed this by invoking logarithmic constraints, $\langle \log k \rangle$.

In summary, the maximization of entropy is a variational prescription for selecting one of many possible models of probability distributions consistent with limited data. In a seminal result that we review here, SJ showed that only the BG entropy or functions with identical maxima ensure that models derived from them satisfy basic logical self-consistency requirements. We apply SJ's approach to derive what joint probability for states of two systems would be required to justify the form of the Tsallis entropy as well as other entropies in selecting model probability distributions consistent with data. We observe that all forms of nonadditive entropy functions require probability rules other than the multiplication rule even when events are independent according to data. We conclude that for modeling nonexponential distributions, such as power laws, nonextensivity should be expressed through the constraints or the $\{q_k\}$, not the entropy. In other words, no structure should be assumed in a distribution function unless it is observed as coupling in the data or originates from the prior distribution on $\{p_k\}$.

We thank the referees for their insightful feedback. S. P. acknowledges support from the Purdue Research Foundation as well as support from his IUPUI Startup. K. A. D. acknowledges support of NIH Grant No. 5R01GM090205-02 and the Laufer Center. K. G. acknowledges support from the Research Corporation for Science Advancement (as a Cottrell Scholar), National Science Foundation (Grant No. 1149992) and PROF grant from the University of Denver.

*Corresponding author.
stevenpresse@gmail.com

[1] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Rev. Mod. Phys. **85**, 1115 (2013).
[2] A. K. Livesey and J. Skilling, Acta Crystallogr. Sect. A **41**, 113 (1985).
[3] J. Skilling, Nature (London) **309**, 748 (1984).
[4] J. Skilling and S. F. Gull, in *Bayesian Maximum Entropy Reconstruction*, Lecture Notes-Monograph Series Vol. 20 (Institute of Mathematical Statistics, Hayward, CA, 1991), pp. 341–367.
[5] S. X. Xie, J. Chem. Phys. **117**, 11 024 (2002).
[6] J. Skilling, in *Maximum-Entropy and Bayesian Methods in Science and Engineering*, edited by G. J. Erickson and C. R. Smith (Kluwer, Dordrecht, 1988), Vol. 1, p. 173.
[7] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[8] E. T. Jaynes, *Probability Theory; the Logic of Science* (Cambridge University Press, Cambridge, England, 2003).
[9] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[10] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).
[11] A. I. Khinchin, *Mathematical Foundations of Information Theory* (Dover, New York, 1957).
[12] C. Hanel and S. Thurner, Eur. Phys. Lett. **93**, 20 006 (2011).
[13] A. Rényi, in *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* (University of California Press, Berkeley, CA, 1960), p. 547.
[14] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **26**, 26 (1980).
[15] P. Jizba and T. Arimitsu, Physica (Amsterdam) **365A**, 76 (2006).
[16] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).
[17] M. Hotta and I. Joichi, Phys. Lett. A **262**, 302 (1999).
[18] C. Tsallis, in *Nonextensive Statistical Mechanics and Its Applications*, edited by S. Abe and Y. Okamoto (Springer, Berlin, 2001), pp. 3–98.
[19] E. M. F. Curado and C. Tsallis, J. Phys. A **24**, L69 (1991).
[20] R. J. V. dos-Santos, J. Math. Phys. (N.Y.) **38**, 4104 (1997).
[21] S. Abe, Phys. Lett. A **271**, 74 (2000).
[22] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, Phys. Rev. Lett. **52**, 1357 (1984).
[23] C. Tsallis, M. Gell-Mann, and Y. Sato, Proc. Natl. Acad. Sci. U.S.A. **102**, 15 377 (2005).
[24] S. Abe, Phys. Rev. E **63**, 061105 (2001).
[25] A. Cho, Science **297**, 1268 (2002).
[26] G. Wilk and Z. Wlodarczyk, Phys. Rev. Lett. **84**, 2770 (2000).
[27] C. Beck, Phys. Rev. Lett. **87**, 180601 (2001).
[28] C. Tsallis, R. S. Mendes, and A. R. Plastino, Physica (Amsterdam) **261A**, 534 (1998).
[29] S. Abe and G. B. Bagci, Phys. Rev. E **71**, 016139 (2005).
[30] S. Abe, Phys. Rev. E **79**, 041116 (2009).
[31] S. Abe, Eur. Phys. Lett. **84**, 60 006 (2008).
[32] R. Hanel, S. Thurner, and C. Tsallis, Eur. Phys. Lett. **85**, 20 005 (2009).
[33] S. Abe, J. Stat. Mech. (2009) P07027.
[34] A. N. Gorban and I. V. Karlin, Phys. Rev. E **67**, 016104 (2003).
[35] B. Mandelbrot, in *Communication Theory*, edited by W. Jackson (Butterworth, Woburn, MA, 1953), pp. 486–502.