Nonadditive entropy maximization is inconsistent with Bayesian updating

Steve Pressé[*]

Department of Physics, IUPUI Indianapolis, Indiana 46202, USA

(Received 30 June 2014; revised manuscript received 18 July 2014; published 24 November 2014)

The maximum entropy method—used to infer probabilistic models from data—is a special case of Bayes's model inference prescription which, in turn, is grounded in basic propositional logic. By contrast to the maximum entropy method, the compatibility of nonadditive entropy maximization with Bayes's model inference prescription has never been established. Here we demonstrate that nonadditive entropy maximization is incompatible with Bayesian updating and discuss the immediate implications of this finding. We focus our attention on special cases as illustrations.

DOI: 10.1103/PhysRevE.90.052149 PACS number(s): 05.20.−y, 02.50.Tt, 89.70.Cf

I. INTRODUCTION

Maximum entropy, MaxEnt, is widely used to infer probabilistic models from limited data [1]. For instance, MaxEnt has been used to determine size distributions of heterogeneous particles in solutions [2], protein folding kinetic rates [3], and pair distance distributions from electron spin resonance [4].

Probability distributions, $p(x)$, inferred using MaxEnt satisfy two key criteria:

Criterion 1: They are consistent with whatever constraints, C, are available from the data.

Criterion 2: They simultaneously maximize Shannon's entropy, $S = -\int dx p(x) \log p(x)$ [5–7].

While Criterion 1 is uncontroversial, Criterion 2 may appear *ad hoc* [e.g., Why not maximize another arbitrary function of $p(x)$?]. Reference [7] quantitatively addresses the fundamental reasons why maximizing the Shannon entropy should be preferred. For example, Criterion 2 assures that joint distributions inferred from MaxEnt models, say, $p(x,y)$, where x and y are model variables, factorize into the product of marginal distributions unless the constraints from the data warrant otherwise.

Criterion 2 imposes properties that probabilistic models should satisfy prior to considering data. For this reason, the Shannon entropy should be related to a Bayesian model prior. And, indeed, the MaxEnt model inference procedure—with the Shannon entropy playing the role of a prior—is a special case of Bayesian inference [8] which, itself, is firmly grounded in basic propositional logic [9].

Here we first review how MaxEnt is related to Bayesian inference and subsequently show that nonadditive entropy maximization cannot be reconciled with Bayes's model inference scheme. In doing so, we will have shown why distributions inferred using nonadditive entropy maximization lack predictive power and, simultaneously, illustrate the power of Bayes's theorem in establishing the merit of an inference method.

II. MAXENT IS A SPECIAL CASE OF BAYESIAN INFERENCE

Bayes's theorem is a principled recipe that can be used to update probability distributions over models, $P(M = p(x))$, when data, D, become available.

For instance, prior to considering data, we have an opinion on how probable a certain model is and this is quantified by the distribution $P(M)$ (called the model prior). The central object of interest in Bayesian analysis is not $P(M)$ but rather the probability of the model once data has actually been gathered, $P(M|D)$. To construct the posterior probability, $P(M|D)$, we update the model prior $P(M)$ by invoking Bayes's theorem,

$$P(M|D) \propto P(D|M)P(M), \tag{1}$$

where $P(D|M)$, a known function, is the probability of the data given the model (called a likelihood function).

MaxEnt is a special case of Bayes's model inference prescription for these reasons: (1) MaxEnt returns one optimal model which we call M^* [as opposed to an entire distribution over models $P(M|D)$]. This model is the one that maximizes $P(M|D)$. Thus, by assumption, the MaxEnt procedure assumes $P(M|D)$ is sharply peaked at its maximum. (2) MaxEnt assumes a special form for the model prior, $P(M)$. To find the form for $P(M)$ dictated by MaxEnt, we relate the objective function commonly maximized in MaxEnt (entropy plus constraints) to Eq. (1),

$$C(D|M) + S(M) \leftrightarrow P(D|M)P(M) \tag{2}$$

from which it can be shown that $P(M) \propto e^S$ [8].

MaxEnt is a versatile inference scheme grounded in Bayesian inference. Indeed, the MaxEnt prescription can infer any type of probability distribution. One common textbook application of MaxEnt is to use an average of a random variable as a constraint, $C = \lambda[\int dx x p(x) - \bar{x}]$, where λ is a Lagrange multiplier. By maximizing the Shannon entropy subject to this constraint, the MaxEnt procedure yields an exponential distribution [1]. Yet the Shannon entropy does not only generate simple exponentials. In one of the earliest and perhaps best-cited nontrivial applications of MaxEnt [10], an entire probability density, whose grayscale represents remnant features of the supernova Cas A, was inferred. This million-pixel two-dimensional distribution is much more complex than any simple distribution having an analytic form (such as an exponential).

III. NONADDITIVE ENTROPIES CANNOT BE RECONCILED WITH BAYESIAN UPDATING

Despite the success of the Shannon entropy prior (e^S) [1], Shannon's entropy is often substituted for the nonadditive

[*] spresse@iupui.edu

entropy [11–13], S_q, which has controversially been conjectured in physics as a generalization of the standard entropy formula [14],

$$S_q = \frac{1}{1-q}\left[\int dx p(x)^q - 1\right]. \tag{3}$$

Here q is a parameter called the "entropic index" and, for concreteness, we take x to be a continuous random variable with range $[0,\infty]$.

Contrary to traditional MaxEnt, we demonstrate that nonadditive entropy maximization cannot be reconciled with Bayesian updating. To show this, we begin by writing a hierarchical Bayesian scheme to explicitly accommodate the entropic index q,

$$P(M,q|D) \propto P(D|M,q)P(M|q)P(q), \tag{4}$$

where we condition the model prior, $P(M|q)$, on the hyperparameter q. Equation (4) is an expanded form of Bayes's theorem, Eq. (1). It is not an approximation. More generally, Eq. (4) treats q as a random variable subject to its own prior distribution rather than treating q as a fixed value. To be clear, a fixed value for q would imply the special case of a delta function prior over q.

Hyperparameters (such as q) are routine in Bayesian modeling. Bayes's theorem does not impose restrictions on the form for $P(q)$. It only strictly requires that $P(q)$ be some prespecified function before considering data (and thus not be an explicit function of D). The Rényi entropy, for instance, also introduces a hyperparameter (usually called α) whose value must also be quantified [15].

As we now detail, the reason that nonadditive entropy maximization is not consistent with Bayesian updating is because $P(q)$ is directly informed by the data.

In fact, as we will see, $P(q)$ will be set to a delta function whose center, q^*, is treated as an adjustable parameter which—in the words of the statistician J. Berger [16]—is "perhaps the most questionable of all the pseudo-Bayes procedures [which] is to write down proper (often conjugate) priors with unspecified parameters, and then to treat these parameters as 'tuning' parameters to be adjusted until the answer 'looks nice'. At the very least, anyone using this technique should clearly explain that this is what was done," adding that "...while these pseudo-Bayesian techniques can be useful as data exploration tools, they should not be confused with formal objective Bayesian analysis, which has very considerable extrinsic justification as a method of analysis."

In the nonadditive entropy literature the value set for q [17], which we will call q^*, depends explicitly on the data. Either q^* is obtained by curve fitting [17] or conjectured to match expected properties of the system that give rise to the data [18,19].

Thus, consistent with the logic of fixing q^* according to some data which we preliminarily call D', we obtain the model posterior by marginalizing $P(M,q|D)$ [given by Eq. (4)] over q,

$$\begin{aligned} P(M|D) &\propto \int dq P(D|M,q)P(M|q)\delta[q - q^*(D')] \\ &= P(D|M,q^*(D'))P(M|q^*(D')), \end{aligned} \tag{5}$$

where D' is the data that determine q^*.

If D and D' have no data in common (i.e., they are disjoint sets), then Eq. (5) is compatible with Bayesian updating. Otherwise, if D and D' do have data in common, Eq. (5) becomes

$$P(M|D) \propto P(D|M,q(D))P(M|q(D)). \tag{6}$$

We will show in the subsection that follows that nonadditive entropy maximization leads to the incorrect form of Bayes's theorem given by Eq. (6).

Equation (6) is incorrect because it treats the very same data, D, on a different footing (i.e., the data inform the prior while, simultaneously, informing the likelihood function).

The dependence of q on the data in Eq. (6) is reminiscent of a method called empirical Bayes [20]. In empirical Bayes, hyperprior parameters (such as q) are fixed by the data and, for this very reason, empirical Bayes is also not free of criticism [16]. A strategy sometimes used in empirical Bayes is to marginalize $P(D|M,q)P(M|q)$ over M to obtain $P(D|q)$. Given $P(D|q)$, point statistics such as moments of this distribution (which depend on q) could be set to their values obtained from data to estimate q.

Probability distributions [i.e., models $M = p(x)$] drawn from Eq. (6) have limited predictive power especially in the limit of small data sets.

Now we turn to a concrete example as an illustration.

A. An illustration using power-law distributions

Power-law distributions arise in the nonadditive entropy framework by maximizing the following objective function:

$$\begin{aligned} \phi_q \equiv{}& \frac{1}{1-q}\left[\int dx p(x)^q - 1\right] \\ &- \alpha\left[\int dx p(x) - 1\right] - \lambda\left[\int dx x p(x) - \bar{x}\right], \end{aligned} \tag{7}$$

which is a nonadditive entropy, S_q, constrained by an average (enforced by Lagrange multiplier λ) and a normalization over the $p(x)$ (enforced by the Lagrange multiplier α).

The model maximizing ϕ_q is the q exponential

$$p_q^*(x) = \frac{\theta}{\bar{x}(\theta-1)}\left[1 + \frac{x}{\bar{x}(\theta-1)}\right]^{-\theta-1}, \tag{8}$$

where, for notational convenience, we have defined $-\theta - 1 \equiv 1/(q-1)$ [21] and, for the sake of clarity, we add a q subscript to all distributions obtained from nonadditive entropy maximization, p_q^*. We note, for completeness, that $p_q^*(x)$ only has a well-defined mean if $1 \geqslant q > 1/2$. For fits to data requiring a q beyond $1 \geqslant q > 1/2$, the form for $p_q^*(x)$ must be altered. This has been accomplished by postulating new definitions for averages in ϕ_q [from $\int dx x p(x)$ to $\int dx x (p(x))^q$], thereby drawing fresh criticism [22] on what is already a controversial model inference scheme [23–26].

So far, the recipe we just followed—in going from ϕ_q to $p_q^*(x)$—can be repitched in the Bayesian framework (problems will arise as we try to parametrize q). In particular, the model we just inferred, $p_q^*(x)$, equivalently follows from the maximization of the following posterior [1]:

$$P(D|M,q)P(M,q) \propto \delta\left[\int dx x p(x) - \bar{x}\right]\delta\left[\int dx p(x) - 1\right]e^{S_q}. \qquad (9)$$

Both procedures, that is, the maximization of Eq. (9) or (7), so far have left q in $p_q^*(x)$ undetermined. Bayes's model inference framework would have enforced that we specify a prior distribution of whichever form (even flat) over q. By contrast, the nonadditive entropy framework [17] requires that q be informed from the data. We now follow the latter recipe to its logical conclusion (the eventual incompatibility with Bayesian updating).

Suppose we take the curve-fitting route to determine q (actually using the alternate route of "deriving" q from system properties would not change our conclusions because, ultimately, q derives from D).

To curve fit q, we follow a standard statistical recipe [21] by first constructing a likelihood function $L(\theta) = \prod_i^n p_q(x_i)$ for n identical independently distributed data points, $D = \{x_i\}$. We then select the value of θ which maximizes the likelihood, θ^*, which is satisfied by the following self-consistent equation:

$$\theta^* = \frac{n}{\sum_i \left\{ \log\left[1 + \frac{x_i}{\bar{x}(\theta^*-1)}\right] - \frac{1}{\theta^*-1}\left[\frac{x_i(\theta^*+1)}{x_i+\bar{x}(\theta^*-1)} - 1\right]\right\}}. \qquad (10)$$

In the large-data-set limit ($n \to \infty$), this estimate should coincide with any exact q "derived" for the system. Equation (10) pinpoints exactly which statistics q^* explicitly depends on. For instance, q^* depends explicitly on $\bar{x}$. Hence, we find that the data D' on which q^* depends [in Eq. (5)] is the same data D which appears in the likelihood function $P(D|M,q^*(D'))$. In doing so, we have illustrated that Eq. (6) holds for even this most basic example. Of course, any q determined from curve fitting—no matter the functional form for $p_q^*(x)$—will clearly depend on the data, D, that make up that curve which, incidentally, also informs the likelihood function.

B. The improbable q exponential

The parametrization of q, as it is currently accomplished following the nonadditive entropy framework, eliminates all predictive power for any $p_q^*(x)$.

Nonetheless, any distribution function $p_q^*(x)$, obtained by maximizing the nonadditive entropy under arbitrary constraints, may be normalizable and everywhere positive. This was certainly true of our special case—the q-exponential given by Eq. (8)—for some q range. Thus any $p_q^*(x)$ can still play the role of a probability distribution function.

Now suppose that we use the very same data to obtain: (1) a MaxEnt distribution $p^*(x)$ and (2) a distribution obtained from nonadditive entropy maximization, $p_q^*(x)$. But let us assume that we do not parametrize the q in $p_q^*(x)$. It is now fair to ask how probable the model $p_q^*(x)$ is as compared to the MaxEnt model $p^*(x)$ for any value of q if the same data are used to inform both distributions.

To compare $p_q^*(x)$ with $p^*(x)$, we consider the ratio of posteriors evaluated at $p_q^*(x)$ and $p^*(x)$,

$$\frac{P(M \equiv p_q^*(x)|D)}{P(M \equiv p^*(x)|D)} = e^{\int dx p^*(x) \log p^*(x) - \int dx p_q^*(x) \log p_q^*(x)}. \qquad (11)$$

In Eq. (11) we note that the ratio of posteriors simplifies to a ratio of priors because—by assumption—the likelihood functions are identical.

By construction, since $p^*(x)$ maximizes the Shannon entropy, then the ratio of posteriors must be less than 1 [or identical to 1 if $p^*(x) = p_q^*(x)$]. Equation (11) is as far as we can go unless we specify what data was used to inform $p_q^*(x)$ and $p^*(x)$.

For concreteness, therefore, we consider the special case where we have constraints on a mean value $\bar{x}$. We consider this special case because it is the simplest and most common in the literature. Of course, any other constraints can be used. For this special case, $p_q^*(x)$ takes the form of Eq. (8) and $p^*(x)$ takes its usual exponential form, $\frac{1}{\bar{x}}e^{-x/\bar{x}}$. We then find

$$\frac{P(M \equiv p_q^*(x)|D)}{P(M \equiv p^*(x)|D)} = \frac{e^{\frac{1-q}{q}}(-1+2q)}{q}, \qquad (12)$$

which is again only valid for $1 \geqslant q > 1/2$ [a requirement that the mean of $p_q^*(x)$ be finite] and reduces to unity in the limit that $q \to 1$, as it should.

Now consider a model where $q \sim 0.5$, say, in Eq. (12). Such a model is substantially less likely to be a valid model—as compared to the MaxEnt distribution under average constraints—and would, normally, be discarded according to the logic of Eq. (12).

In fact, this logic generalizes to all models with $q \neq 1$ and the exercise can be repeated for any distributions (not just power laws and exponentials) obtained from any data constraints.

IV. CONCLUSION

Historically, the nonadditive or "Tsallis" entropy emerged as a early effort to infer exotic distributions at a time when principled inference techniques were largely unknown to the broader physics community. While, at first glance, the generalized form for the Shannon entropy [Eq. (3)] might seem plausible, careful investigation over the past decade has shed light on the perplexing inconsistencies to which this peculiar generalization gives rise [22,25,26].

For instance, the models obtained by maximizing nonadditive entropies are suboptimal as compared to those obtained by maximizing the Shannon entropy because—as we have shown [26]—they introduce biases in the inferred probabilities where none are warranted by the data. In fact, only $q = 1$ (which is the limit in which nonadditive entropies become the Shannon entropy) assures that the inferred distribution does not introduce spurious correlations between model variables where none are otherwise warranted. These spurious correlations are made explicit in Eq. (13) of Ref. [26].

Here we show definitively that invoking nonadditive entropy maximization to infer models yields probability distributions of no predictive value because they are obtained in a way that is incompatible with Bayesian updating.

By analogy, suppose we were to alter Maxwell's equations by changing the power of their spatial and temporal derivatives to model electromagnetic wave propagation through some exotic materials. While the fit to some data sets might improve, we would—just as a start—lose Lorentz invariance and overturn the definitions of dielectric and magnetic permeabilities to accommodate the changes in units of our fields.

There is intrinsic value in searching for generalizations of basic physical principles to tackle complex problems presented by biology and other systems out of equilibrium. However, basic foundational principles—such as Bayesian updating rules—must be preserved.

The emerging perspective is that Shannon's entropy provides a way of finding the most probable model consistent with laws of inference [8] and it is not restricted to equilibrium or other simple systems [1] as the breadth of its generality is demonstrated by Shore and Johnson [7].

ACKNOWLEDGMENTS

S.P. thanks Ken Dill, Julian Lee, Kingshuk Ghosh, and George Judge for helpful discussions. S.P. also acknowledges support from the NSF (Grant No. MCB-1412259).

[1] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Rev. Mod. Phys. **85**, 1115 (2013).
[2] P. Sengupta, K. Garai, J. Balaji, N. Periasamy, and S. Maiti, Biophys. J. **84**, 1977 (2003).
[3] P. J. Steinbach, R. Ionescu, and C. R. Matthews, Biophys. J. **82**, 2244 (2002).
[4] Y. W. Chiang, P. P. Borbat, and J. H. Freed, J. Magn. Res. **177**, 184 (2005).
[5] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).
[6] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[7] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **26**, 26 (1980).
[8] J. Skilling and S. F. Gull, Lect. Notes Monogr. Ser. **20**, 341 (1991).
[9] E. T. Jaynes, *Probability Theory; The Logic of Science* (Cambridge University Press, Cambridge, 2003).
[10] J. Skilling and R. K. Bryan, Mon. Not. R. Astron. Soc. **211**, 111 (1984).
[11] J. Havrda and F. Charvát, Kybernetika **3**, 30 (1967).
[12] J. Burbea and C. R. Rao, IEEE Trans. Inf. Theory **IT-28**, 489 (1982).
[13] J. Burbea and C. R. Rao, IEEE Trans. Inf. Theory **IT-28**, 961 (1982).
[14] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).
[15] A. G. Bashkirov, Phys. Rev. Lett. **93**, 130601 (2004).
[16] J. Berger, Bayesian Anal. **1**, 385 (2006).
[17] C. Tsallis, Physica A **221**, 277 (1995).
[18] P. Douglas, S. Bergamini, and F. Renzoni, Phys. Rev. Lett. **96**, 110601 (2006).
[19] E. Lutz, Phys. Rev. A **67**, 051402 (2003).
[20] G. Casella, The American Statistician **39**, 83 (1985).
[21] C. R. Shalizi, arXiv:math/0701854v2 (2007).
[22] S. Abe, Phys. Lett. A **275**, 250 (2000).
[23] J. Cartwright, Phys. World 31 (2014).
[24] A. Cho, Science **297**, 1268 (2014).
[25] M. Nauenberg, Phys. Rev. E **67**, 036114 (2003).
[26] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Phys. Rev. Lett. **111**, 180604 (2013).