Beyond Boltzmann-Gibbs statistics: Maximum entropy hyperensembles out of equilibrium

Gavin E. Crooks[*]

Physical Bioscience Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

(Received 6 March 2006; revised manuscript received 2 March 2007; published 27 April 2007)

What is the best description that we can construct of a thermodynamic system that is not in equilibrium, given only one, or a few, extra parameters over and above those needed for a description of the same system at equilibrium? Here, we argue the most appropriate additional parameter is the nonequilibrium entropy of the system. Moreover, we should not attempt to estimate the probability distribution of the system directly, but rather the metaprobability (or hyperensemble) that the system is described by a particular probability distribution. The result is an entropic distribution with two parameters, one a nonequilibrium temperature, and the other a measure of distance from equilibrium. This dispersion parameter smoothly interpolates between certainty of a canonical distribution at equilibrium and great uncertainty as to the probability distribution as we move away from equilibrium. We deduce that, in general, large, rare fluctuations become far more common as we move away from equilibrium.

Consider a gas confined to a piston, as illustrated in Fig. 1. The realization on the left was sampled from thermal equilibrium with a fixed plunger. To describe the probability of every single possible configuration of the particles we only need to know the Hamiltonian of the system and the temperature of the environment [1,2]. On the other hand, the system on the right has been sampled from a nonequilibrium ensemble. Although the Hamiltonian is the same, the plunger has recently been in violent motion and this perturbation has driven the ensemble away from equilibrium. To describe the configurational probability we now need to know the entire past history of perturbations that the system has undergone. The dynamics and historical details matter.

This example illustrates the essential difficulty we face when trying to directly extend equilibrium statistical mechanics out of equilibrium. There is only one ensemble that can describe a given system in thermal equilibrium, but there are a multitude of ways that the same system can be out of equilibrium. That the equilibrium entropy is maximized (given the available constraints, such as the mean energy) is a strong condition that uniquely determines the probability distribution. However, let us take a step back, and reflect that statistical mechanics itself is designed to circumvent a similar difficulty. In classical mechanics we typically assume that we know the exact microstate of the system. However, in statistical mechanics we recognize that such a detailed description is frequently neither possible nor desirable. A few bulk measurements or parameters do not provide nearly enough information to fix the microstate. Instead we content ourselves with calculating the probability that the system occupies a particular microstate. To ask what the state of the system *is*, rather than what it *could be*, is to ask an unnecessarily difficult question.

Out-of-equilibrium we essentially face the same problem, compounded. Clearly we cannot obtain enough information from a few measurements to determine the microscopic state

of the system, but if the system is out-of-equilibrium then a few parameters or measurements are not sufficient (in general) to determine the ensemble either. Therefore, perhaps the correct approach is not to try to determine what the probability distribution of the system is, but instead attempt to determine what the probabilities could be. In other words, instead of thinking about an ensemble of systems, we instead envisage an ensemble of ensembles, a "hyperensemble" (Fig. 2), where each member of the hyperensemble has the same instantaneous Hamiltonian, but is described by a different probability distribution. We seek a generic description of the typical nonequilibrium ensemble given a few parameters or measurements that describe the average behavior of the hyperensemble.

We can think of this approach as an extension of the method of the most probable distribution [1,3,4]. If the canonical distribution is the most probable in equilibrium, we may reasonable ask what are the highly plausible distributions as we move away from equilibrium. This basic idea of estimating the probability of a probability density (a "metaprobability") is often used in Bayesian statistics, especially when the available data is too sparse to reliable estimate the probability directly [5–7]. Reference [6] contains a lucid description of this procedure in the context of amino acid sequence profiles. We have borrowed the hyperprefix from Bayesian statistics, were it is usual to talk about hyperpriors (a prior distribution of a prior distribution) and associated hyperparameters.

With this insight, we can move beyond the standard canonical ensemble by changing the question. Instead of trying to find the probability distribution θ of the system directly, we instead estimate the metaprobability $P(\theta)$, the probability
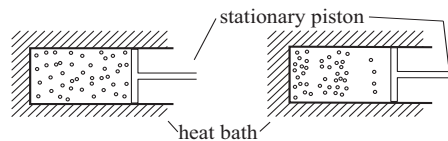

FIG. 1. Schematic realizations of a gas confined to a piston in and out of equilibrium.

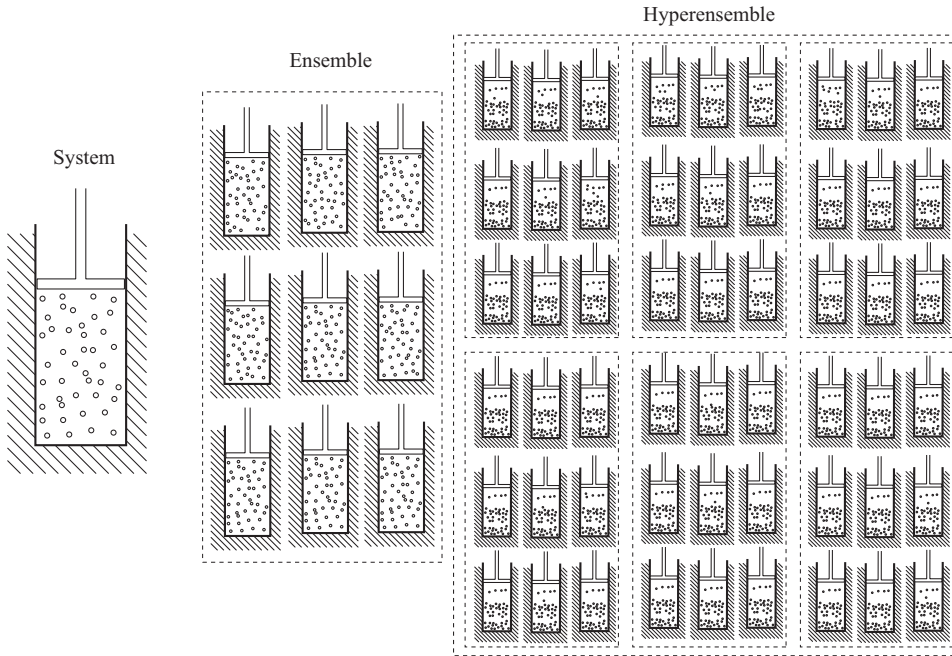*Electronic address: gecrooks@lbl.gov; URL: http://bespoke.lbl.gov/

FIG. 2. Schematic illustration of a single system, an ensemble of systems and a hyperensemble, an ensemble of ensembles.

of the microstate probability distribution. We proceed analogously to the maximum entropy derivation of equilibrium statistical mechanics [2,8]. We consider a physical system with a set of states, each characterized by an energy E_i. We will find the probability distribution of ensembles $P(\theta)$ that maximizes the entropy $\mathcal{H}$ of the hyperensemble

$$\mathcal{H}(P(\theta)) = -\int P(\theta) \ln \frac{P(\theta)}{m(\theta)} d\theta, \qquad (1)$$

while maintaining certain appropriate constraints. Here, θ is the positive vector $\{\theta_1, \theta_2, \ldots, \theta_K\}$ and integration is performed over normalized probabilities:

$$d\theta = \delta\left(\sum_{i=1}^{K} \theta_i - 1\right) d\theta_1 d\theta_2 \cdots d\theta_K.$$

The distribution of distributions $P(\theta)$ is a normalized map of θ to real numbers $[\int P(\theta) d\theta = 1]$.

The entropy is measured relative to the distribution $m(\theta)$. This distribution acts as a prior for θ and consequentially we should set $m(\theta)$ to the most uninformative distribution consistent with the prior data [5,7]. In the current case, we only know that there are K accessible states and that we have no reason to favor any state over any other state. Consequentially, the appropriate prior is the uniform distribution $m(\theta)$ =const.

The trick to maximum entropy methods is finding the appropriate constraints, since with an arbitrary choice of constraint and prior practically any answer can be manufactured. To avoid this trap, we seek a minimal set of physically and mathematically reasonable parameters. Clearly, the hyperensemble must be normalized,

$$1 = \int P(\theta) d\theta. \qquad (2)$$

And, by analog with the canonical ensemble, we should constrain the mean energy of the ensemble of ensembles

$$\langle\langle E \rangle\rangle = \int P(\theta) \left[\sum_i \theta_i E_i\right] d\theta, \qquad (3)$$

where E_i is the energy of state i.

Thus far, we have only incorporated the same information and constraints that lead to the canonical ensemble, namely, the density of energy states, normalization, and mean energy. In addition we require a measure of how far the system is from equilibrium. After all, the quintessential feature of nonequilibrium systems is that they are not in equilibrium. What is the most appropriate measure? If the system were in equilibrium, then the entropy would be maximized given the constraints. It follows that out-of-equilibrium the entropy of the ensemble is not maximized, and moreover, the entropy cannot be determined with any certainty from a measurement of the mean energy alone. Therefore, the entropy itself can be used as an additional, physically relevant constraint:

$$\langle S \rangle = \int P(\theta) \left[-\sum_i \theta_i \ln \theta_i\right] d\theta. \qquad (4)$$

To summarize, we will maximize the entropy of the hyperensemble [Eq. (1)] subject to normalization, the mean energy and the mean ensemble entropy [Eq. (2)–(4)]. The solution to this problem is found by introducing Lagrange multiplies $\{\lambda\}$ and then applying the calculus of variation in the usual way:

$$P(\theta) = e^{-\lambda_0 - \lambda_1 \langle E \rangle - \lambda_2 S(\theta)}. \qquad (5)$$

Some manipulation will illuminate the significance of this expression. Let us rewrite with $\lambda_1 = \lambda\beta$ and $\lambda_2 = \lambda$:
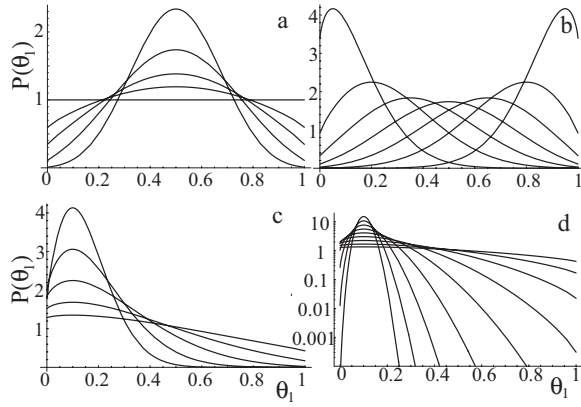
FIG. 3. The entropic distribution [Eq. (8)] over two states. (a) Reference distribution $\rho=(0.5,0.5)$, $\lambda=0,1,2,4,8$ (broad to peaked). (b) $\lambda=4$, $\rho_1=0.05,0.20,0.35,0.5,0.65,0.80,0.95$ (left to right). (c) $\rho=(0.1,0.9)$, $\lambda=0.5,1,2,4,8$. (d) Same ρ, log scale, $\lambda=0.5,1,2,4,8,16,32,64,128$. Note that the reference distribution controls the mode and that as the dispersion parameter λ approaches 0 the distributions become broader and the mean moves towards $\frac{1}{2}$.

$$P(\theta) \propto \exp\left(-\beta\lambda\sum_i \theta_i E_i + \lambda\sum_i \theta_i \ln \theta_i\right). \qquad (6)$$

The parameter β has units of entropy per unit of energy and is effectively an inverse temperature. Therefore, we can naturally introduce a canonical ensemble with the same effective temperature

$$\rho_i = \frac{1}{Q(\beta)} \exp(-\beta E_i) \qquad (7)$$

and rewrite the maximum entropy hyperensemble as

$$P(\theta) = \frac{1}{\mathcal{L}(\beta,\lambda)} \exp\left(-\lambda\sum_i \theta_i \ln \frac{\theta_i}{\rho_i}\right), \qquad (8)$$

where $\mathcal{L}$ is a normalization constant. It is now evident that our hyperensemble has the functional form of the entropic distribution, a probability of probabilities that occasionally occurs in Bayesian statistics [9–15]. This same functional form also appears as the asymptotic limit of the multinomial distribution with large sample sizes [16] and in large deviation theory [16,17].

The entropic distribution over a binary state space is illustrated in Fig. 3 and with a Gaussian reference (e.g., a particle in a harmonic potential) in Fig. 4. We see that as λ decreases the dispersion of the probability distributions increases, the mean distribution moves away from the canonical distribution, the average probability of rare states increases, and the probability of common states decreases to compensate. Moreover, in Fig. 4 we see that λ controls a crossover in behavior; if $\rho>\lambda^{-1}$ then the uncertainty in θ and the bias away from equilibrium are relatively small, whereas for rare states, $\rho<\lambda^{-1}$, the perturbation away from equilibrium are large. Therefore, the generic, predicted behavior is
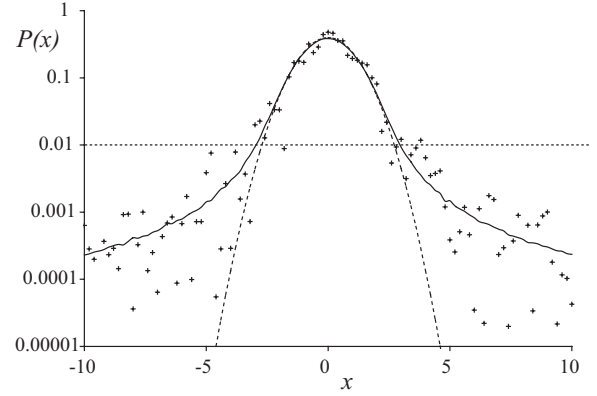

FIG. 4. The entropic distribution [Eq. (8)] with a Gaussian reference distribution (zero mean, unit variance) and dispersion $\lambda=100$. The dashed line is the reference ρ, the points are a single realization of θ and the solid line is the mean distribution $\langle\theta\rangle$ (sampled using a discretized distribution and Monte Carlo Gibbs sampling [21]). Note that the variation of θ away from the reference is relatively large for intrinsically rare states $\rho<1/\lambda$. The crossover in behavior is indicated by the horizontal dashed line.

that rare events typically (but not necessarily) become far more common as the condition of thermal equilibrium is relaxed.

We can deduce some important properties of the hyperensemble by noting that the function in the exponential of Eq. (8) is the relative entropy of θ to the reference canonical distribution ρ [16]:

$$D(\theta\|\rho) = \sum_i \theta_i \ln \frac{\theta_i}{\rho_i}. \qquad (9)$$

This is a natural measure of how distinguishable one distribution is from another. Since the relative entropy is zero if the distributions are identical, and positive if they are not, it immediately follows that the mode of the entropic distribution is located at the reference ρ. In other words, the single most probable distribution of the hyperensemble is a canonical distribution controlled by the effective temperature β, and the dispersion of the hyperensemble about that mode is controlled by the inverse scale parameter λ. If λ is large the hyperensemble collapses to a single point at the mode and we recover the canonical ensemble of equilibrium statistical mechanics. It follows that the reference temperature is numerically equal to the conventional temperature of the same system with the same mean energy at thermal equilibrium. As λ decreases the dispersion increases and typical distributions differ significantly from the reference, until at $\lambda=0$ every distribution is equally likely.

Another way of looking at the canonical hyperensemble is to note that the relative entropy of θ to a canonical reference ρ can be interpreted as a generalized free energy difference [18]

$$D(\theta\|\rho) = \beta F(\theta) - \beta F(\rho),$$

$$\beta F(p) = -\sum_i p_i \ln p_i + \beta \sum_i p_i E_i. \qquad (10)$$

Since ρ is canonical $F(\rho) = S/\beta - \langle E \rangle$ is the Helmholtz free energy, whereas $F(\theta)$ can be interpreted as a generalized, noncanonical free energy. Using these definitions, the canonical hyperensemble can be written as

$$P(\theta) = \frac{1}{\mathcal{L}(\beta,\lambda)} \exp\{-\lambda\beta[F(\theta) - F(\rho)]\}. \qquad (11)$$

The physical picture is that near thermal equilibrium the ensemble that maximizes the free energy dominates the hyperensemble. As we move away from equilibrium the free energy is no longer necessarily maximized. Rather the probability of obtaining a particular ensemble out of equilibrium is determined by the generalized free energy difference between that ensemble and the reference canonical ensemble. This expression is pleasingly reminiscent of the thermodynamic fluctuation representation of standard statistical mechanics [19], except we are now looking at fluctuations in ensemble rather than state.

We can also derive the entropic hyperensemble by directly constraining the mean relative entropy $\langle D(\theta \| \rho) \rangle$. From the viewpoint of information theory, this is the average penalty for encoding states of the system, assuming the they are drawn from the reference distribution ρ rather than the true distributions [16]. This measure is similar to the Jensen-Shannon divergence $\langle D(\theta \| \langle \theta \rangle) \rangle$ [20], except that the reference distribution is the mode, rather than the mean of θ.

The entropic distribution is not particularly amenable to analysis. To proceed further, we adopt a simple approximation. We note that $\theta \approx \rho$ if λ is large, and that $D(\rho \| \theta) = D(\theta \| \rho) + O[\sum_i (\theta_i - \rho_i)^3]$. For small λ the deviations are large and this approximation is inaccurate. Therefore,

$$P(\theta) = \frac{1}{\mathcal{L}(\beta,\lambda)} e^{-\lambda D(\theta\|\rho)} \approx \frac{1}{\mathcal{L}(\beta,\lambda)} e^{-\lambda D(\rho\|\theta)} (\text{given that } \theta \approx \rho)$$

$$\propto \prod_{i=1}^{K} \theta^{\alpha_i - 1}, \quad \alpha_i = \lambda\rho_i + 1.$$

This is a Dirichlet distribution. The mode, mean, covariance matrix are [7]

$$\text{mode } (\theta_i) = \rho_i,$$

$$\langle \theta_i \rangle = \frac{\lambda\rho_i + 1}{\lambda + K} \sim \rho_i + O\left(\frac{1}{\lambda}\right),$$

$$\sigma_{ij}^2 = \begin{cases} \dfrac{(\lambda\rho_i + 1)(\lambda + K - \lambda\rho_i - 1)}{(\lambda + K)^2(\lambda + K + 1)} \sim O\left(\dfrac{1}{\lambda}\right), & i = j, \\[2ex] -\dfrac{(\lambda\rho_i + 1)(\lambda\rho_j + 1)}{(\lambda + K)^2(\lambda + K + 1)} \sim O\left(\dfrac{1}{\lambda}\right), & i \neq j. \end{cases}$$

The mode is unaltered by the approximation. As λ decreases the mean distribution moves away from the canonical distribution as $O(\lambda^{-1})$ and the dispersion of the probability distributions increases. If $\rho > \lambda^{-1}$ then the uncertainty in θ, and the

bias away from equilibrium are relatively small. For rare states, any $\rho < \lambda^{-1}$, the perturbation are large. However, large perturbations also invalidate the approximation.

Currently, various modifications or extensions of Boltzmann-Gibbs statistics are being investigated, including Tsallis statistics (which modifies the entropic function) [22] and maximum entropy production (which modifies the constraints) [23]. Perhaps the most similar approach to the present work is superstatistics [24–28], the central idea of which is that a system may be locally in equilibrium (either in time or space), but globally out of equilibrium. Therefore, the system as a whole can be described by a *mixture* of canonical ensembles, each with a different local temperature. In contrast, the components of the maximum entropy hyperensemble are not required to be canonical.

The essentially difficulty with superstatistics is that the distribution of effective temperatures is unconstrained. It is therefore interesting to ask what distribution of local temperature would maximize the hyperentropy given that the members of the hyperensemble are canonical? Since the result will depend on the density of states, let us explore a simple, but important, special case, a collection of harmonic oscillators. The partition function is $Q(\beta) = \beta^{-c}$ and therefore the mean energy scales as $\langle E \rangle = c/\beta$, where the constant "$c$" is proportional to the size of the system. The prior becomes $m(T) \propto 1/T$ [5]. Plugging these relations into Eq. (8) we find

$$P(T) \propto \left(\frac{T}{T^\circ}\right)^{c\lambda - 1} e^{-c\lambda T/T^\circ}, \qquad (12)$$

where T is the effective local temperature and $T^\circ = 1/\beta$ is the reference temperature. Here, with the hyperensemble approach we predict that if the system is linear and locally in equilibrium, then the temperature fluctuations follow a gamma distribution [27,29,30] with mean T° and standard deviation $T^\circ/\sqrt{c\lambda}$. If the temperature fluctuations are not gamma distributed, then either the system is not linear, not in local equilibrium, or we have failed to incorporate some important, pertinent information about the system [5].

It is worth noting that we would have obtained very different results if we had chosen different constraints. In particular, if we maximize the hyperentropy given the mean relative entropy of the reference ρ to the ensemble θ, $\langle D(\rho \| \theta) \rangle$, then we obtain a Dirichlet distribution. This in turn leads to the prediction that the local temperature of a linear system follows an *inverse* gamma distribution, which is known to be equivalent to the nonextensive statistics of Tsallis [22,24,25,31]. This is an intriguing connection, but in contrast to the constraints on the mean entropy and energy that leads to the entropic distribution, the constraint on $\langle D(\rho \| \theta) \rangle$ does not have any immediately obvious deep physical or information theoretic significance.

In this paper, I have argued that a natural way of moving beyond equilibrium Boltzmann-Gibbs statistics is to change the question: Instead of trying to determine what the probability distribution of a system *is*, we instead ask what the probability distribution *could be*. We seek an ensemble of ensembles that captures the generic properties of matter generically out of equilibrium. The solution to this problem is

found by maximizing the entropy of the hyperensemble, given the mean energy and mean ensemble entropy. This yields a physically plausible description of fluctuations away from equilibrium, a natural definition of temperature out of equilibrium, a natural measure of distance away from equilibrium, and the intuitively plausible prediction that rare events typically become far more common as a system moves away from thermal equilibrium.

It remains uncertain to what extent the predictions of the canonical hyperensemble can be applied to a single experiment. At the very least, this approach tells us how uncertain we should be about thermodynamic predications applied to nonequilibrium systems. Alternatively, we can adopt the attitude of Jaynes [5], that if a nonequilibrium ensemble does not follow the entropic distribution, then we have simply failed to incorporate some important, pertinent information about the system. But perhaps certain nonequilibrium systems will self-average, in much the same way that large, thermodynamic systems self-average. Then the predictions of the canonical hyperensemble and the behavior of the system will be independent of the details of how the system is driven from equilibrium. An obvious candidate for such behavior is fully developed turbulence.

[1] L. W. Boltzmann, Wiener Berichte **63**, 397 (1871).

[2] J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (Yale, New Haven, 1902).

[3] E. Schrödinger, *Statistical Thermodynamics* (Cambridge University Press, Cambridge, 1952).

[4] D. A. McQuarrie, *Statistical Mechanics*, 2nd ed. (University Science Books, New York, 2000).

[5] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).

[6] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, Cambridge, 1998).

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. (Chapman & Hall/CRC, New York, 2004).

[8] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[9] J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Kluwer, Dordrecht, 1989), pp. 45–52.

[10] J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by P. F. Fougère (Kluwer, Dordrecht, 1990), pp. 341–350.

[11] C. C. Rodriguez, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Kluwer, Dordrecht, 1989), pp. 415–422.

[12] M. Brand, in *Artificial Intelligence and Statistics*, edited by D. Heckerman and C. Whittaker (Morgan Kaufmann, San Francisco, 1999), Vol. 7.

[13] M. Brand, Neural Comput. **11**, 1155 (1999b).

[14] A. Caticha, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, edited by A. Mohammad-Djafari (Springer, New York, 2001), p. 94.

[15] A. Caticha and R. Preuss, Phys. Rev. E **70**, 046127 (2004).

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[17] R. S. Ellis, Physica D **133**, 106 (1999).

[18] H. Qian, Phys. Rev. E **63**, 042103 (2001).

[19] H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. (Wiley, New York, 1985).

[20] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).

[21] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 (1984).

[22] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).

[23] R. C. Dewar, J. Phys. A **36**, 631 (2003).

[24] C. Beck and E. G. D. Cohen, Physica A **322**, 267 (2003).

[25] C. Tsallis and A. M. C. Souza, Phys. Rev. E **67**, 026106 (2003).

[26] C. Beck, E. G. D. Cohen, and H. L. Swinney, Phys. Rev. E **72**, 056133 (2005).

[27] H. Touchette and C. Beck, Phys. Rev. E **71**, 016131 (2005).

[28] F. Sattin, Eur. Phys. J. B **49**, 219 (2006).

[29] H. Touchette, in *Nonextensive Entropy-Interdisciplinary Applications* (Oxford University Press, Oxford, 2002), p. 159.

[30] F. Sattin and L. Salasnich, Phys. Rev. E **65**, 035106(R) (2002).

[31] G. Wilk and Z. Włodarczyk, Phys. Rev. Lett. **84**, 2770 (2000).