Principles of maximum entropy and maximum caliber in statistical physics

Steve Pressé*

Department of Physics, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana 46202, USA

Kingshuk Ghosh

Department of Physics and Astronomy, University of Denver, Denver, Colorado 80208, USA

Julian Lee

Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea

Ken A. Dill

Laufer Center for Physical and Quantitative Biology and Departments of Physics and Chemistry, Stony Brook University, New York, New York 11794, USA

(published 16 July 2013)

The variational principles called *maximum entropy* (MaxEnt) and *maximum caliber* (MaxCal) are reviewed. MaxEnt originated in the statistical physics of Boltzmann and Gibbs, as a theoretical tool for predicting the equilibrium states of thermal systems. Later, entropy maximization was also applied to matters of information, signal transmission, and image reconstruction. Recently, since the work of Shore and Johnson, MaxEnt has been regarded as a principle that is broader than either physics or information alone. MaxEnt is a procedure that ensures that inferences drawn from stochastic data satisfy basic self-consistency requirements. The different historical justifications for the entropy $S = -\sum_i p_i \log p_i$ and its corresponding variational principles are reviewed. As an illustration of the broadening purview of maximum entropy principles, maximum caliber, which is path entropy maximization applied to the trajectories of dynamical systems, is also reviewed. Examples are given in which maximum caliber is used to interpret dynamical fluctuations in biology and on the nanoscale, in single-molecule and few-particle systems such as molecular motors, chemical reactions, biological feedback circuits, and diffusion in microfluidics devices.

DOI: 10.1103/RevModPhys.85.1115 PACS numbers: 82.20.Pm, 05.40.−a, 89.70.Cf, 02.50.Tt

CONTENTS

*spresse@iupui.edu.

I. INTRODUCTION

The basic principles of statistical physics with applications to physics and chemistry are given in standard texts (Tolman, 1938; Landau and Lifshitz, 1951; Hill, 1956; de Groot and Mazur, 1962; Balescu, 1975; Chandler, 1987; Pathria, 1996; McQuarrie, 2000; Dill and Bromberg, 2011). Even so, in recent years, statistical physics developed in important new ways. As its foundations became more rigorous and better understood, it is now clear why entropy maximization principles (MaxEnt) also apply beyond problems of equilibrium statistical physics to other areas of science and technology (Ben-Naim, 1985; Denbigh and Denbigh, 1985). Entropy maximization principles provide a framework for understanding dynamics that is as sound as its foundations of equilibria. We refer to *path entropy maximization* principles applied to dynamics as *maximum caliber* (MaxCal) (Jaynes and Haken, 1985). Entropy maximization is poised to play a role in interpreting a growing number of new experiments on single molecules or few-particle systems, particularly in biophysics and nanoscience (Hamill *et al.*, 1981; Methfessel *et al.*, 1986; Livesey and Brochon, 1987; Siemiarczuk, Wagner, and Ware, 1990; Steinbach *et al.*, 1992; Hille, 1994; Schnitzer and Block, 1995; Lu, Xun, and Xie, 1998; Liphardt *et al.*, 2001; Elowitz *et al.*, 2002; Yang and Xie, 2002; Yang *et al.*, 2003; Rhoades *et al.*, 2004; Witkoskie and Cao, 2004, 2008; Cecconi *et al.*, 2005; Lezon *et al.*, 2006; Schneidman *et al.*, 2006; Huang *et al.*, 2007; Vergassola, Villermaux, and Shraiman, 2007; Sanchez and Kondev, 2008; Shahrezaei and Swain, 2008; Southworth and Agard, 2008; Moffitt *et al.*, 2009; Tkacik, Walczak, and Bialek, 2009; Eldar and Elowitz, 2010; Walczak, Tkacik, and Bialek, 2010; Yu *et al.*, 2010; Bustamante, Cheng, and Mejia, 2011).

In some cases, MaxEnt and MaxCal differ from more traditional methods of model making (Gillespie, 1977; van Kampen, 1981; Arkin and Ross, 1995; Qin, Auerbach, and Sachs, 1997, 2000; Rao and Arkin, 2003; Ross, 2003; Paulsson, 2004, 2005; Bratsun *et al.*, 2005; El Samad *et al.*, 2005; Flomenbom, Klafter, and Szabo, 2005; Kaern *et al.*, 2005; Kou *et al.*, 2005; Milescu, Akk, and Sachs, 2005; Samoilov, Plyasunov, and Arkin, 2005; Warren and ten Wolde, 2005; Flomenbom and Silbey, 2006; Lipshtat *et al.*, 2006; McKinney, Joo, and Ha, 2006; Schultz *et al.*, 2007; Cao and Silbey, 2008; Raj and van Oudenaarden, 2008; Wang, Xu, and Wang, 2008; Çağatay *et al.*, 2009; Munsky, Trinh, and Khammash, 2009). Rather than assuming a model and adjusting parameters to fit data, MaxEnt and MaxCal are often used in the reverse direction, to infer models more directly from the data themselves.

We first review the history of thought about the foundations of MaxEnt in statistical physics. It has been known for more than a century how to apply the methods of equilibrium statistical physics, namely, how to find equilibrium states on the basis of maximizing entropy or minimizing free energy and how to apply the Boltzmann distribution law. There have been changing viewpoints on how to justify those methods. Such foundations are important because they determine the extent of applicability of the method. We describe three main mileposts along the way: (1) Boltzmann's maximum-multiplicity justification and Gibbs' ensemble method for predicting the equilibrium properties of gases; (2) Jaynes' formulation, based on Shannon's information theory, in which statistical physics is regarded as a matter of making predictions from limited data by assuming maximal ignorance about the unknown degrees of freedom; and (3) more recently, the formulation of Shore and Johnson that views the maximization of entropy in a much broader light, as a fundamental requirement for ensuring that inferences drawn from data satisfy basic self-consistency requirements of probabilities.

II. A BRIEF HISTORY OF MAXIMUM ENTROPY

Statistical physics originated in the mid-1800s in an effort to understand the gas laws. The gas laws were crucial for understanding how to convert heat to motive force during the industrial revolution (Brush, 1975, 1976, 1983). Statistical physics is the story of entropy and its usage as a variational principle for making predictions (Clausius, 1850a, 1850b). In 1865, Clausius coined the term *entropy* to refer to the quantity q/T (Brush, 1975), where q is the heat and T is temperature. Entropy arose as a key predictor of equilibria when used in conjunction with its maximization principle, called the second law. The average velocity of gas molecules at equilibrium in a container is $\langle \vec{v} \rangle = 0$. However, the *mean-square velocity* is not zero; early theoretical developments showed it to be $\langle \vec{v}^2 \rangle = 3k_BT/m$, where k_B is Boltzmann's constant, T is temperature, and m is the mass of the gas molecule. This expression was central to formulating the kinetic theory of gases. The implication of this nonzero variance is that understanding gas behavior requires *distribution functions*, not just average quantities. In 1858, Clausius derived the mean free paths of gas molecules using probabilistic arguments based on the law of the distribution of errors (Brush, 1976). Maxwell made more quantitative arguments in the 1860s, predicting that such velocity distributions were Gaussian (Brush, 1975). Boltzmann generalized this further, treating gases in the presence of potentials (Lindley, 2001), and deriving what is now called the Maxwell-Boltzmann distribution, thus establishing the basis for the kinetic theory of gases, a major success for theoretical physics at the time.

Boltzmann's justification of the Maxwell-Boltzmann distribution was based on the following ideas. Gas particles occupy small volumetric cells $i = 1, 2, 3, \ldots, s$ of phase space, with occupation numbers n_i. Particle number is conserved so $\sum_i n_i = N$, where the constant N is the number of gas molecules. The number of ways W that any particular distribution $\{n_i\}$ of particles will fall within their respective volumetric cells in phase space is given by the multinomial formula

$$W = N!/(n_1! \ldots n_s!). \tag{1}$$

Boltzmann recognized that the entropy S was proportional to the logarithm $\log W$ of the multiplicity. Taking the

logarithm of W and approximating the factorial for large N using Stirling's formula gives

$$\log W = \log N! - \sum_i \log n_i! \approx -N\sum_i (n_i/N)\log(n_i/N)$$
$$= -N\sum_i p_i \log p_i, \tag{2}$$

where $p_i = n_i/N$ is taken to be the probability that a particle is in cell i, provided N is sufficiently large. Boltzmann then asserted that the occupation probabilities $\{p_i\}$ of the most probable state at equilibrium are those that maximize the entropy, $S = -k_B\sum_i p_i \log p_i \propto \log W$, and which also satisfy two constraints on the total particle number N, and on the average energy (per particle) $\bar{\varepsilon}$,

$$\sum_i p_i = 1; \qquad \sum_i p_i \varepsilon_i = \bar{\varepsilon}. \tag{3}$$

The proportionality constant k_B, Boltzmann's constant, sets the units of S. The state of equilibrium is computed by maximizing the entropy subject to the constraints, using the variational function

$$S(\{p_i\})/k_B - \beta\left(\sum_i p_i \varepsilon_i - \bar{\varepsilon}\right) - \alpha\left(\sum_i p_i - 1\right), \tag{4}$$

where β and α are Lagrange multipliers. This variational function has $s+2$ unknowns: s different p_i's and two Lagrange multipliers. Variation of Eq. (4) with respect to each p_i and the Lagrange multipliers uniquely determines the $s+2$ unknowns. The $\{p_i\}$ values that maximize the variational function are

$$p_i^* = \frac{e^{-\beta\varepsilon_i}}{\sum_k e^{-\beta\varepsilon_k}}. \tag{5}$$

This is called the Boltzmann distribution. The $\{p_i^*\}$ values predict the most probable occupation probabilities at equilibrium. Boltzmann's approach did not require more detail, such as knowledge of the individual trajectories of the particles. His reasoning provided the logic for solving a highly underdetermined problem. He was able to assign a value to the large number of unknowns, the $\{p_i^*\}$ and the Lagrange multipliers, from two simple constraints. Boltzmann's reasoning has been described as a "superefficient" way to capture the essential mathematical ingredients (Jaynes, Levine, and Tribus, 1979): "Whether by luck or inspiration, he [Boltzmann] put into his equations only the dynamical information that happened to be relevant to the questions he was asking."

However, Boltzmann's neglect of the details of individual trajectories was controversial at the time. Justifying his predictions without using a system's dynamics appeared to require the *ergodic hypothesis*, the assertion that the time average of a property taken over its dynamical trajectory equals the equilibrium average of that property over its equilibrium ensemble. Much work followed and continues to explore the relevance of treating the detailed dynamics. Boltzmann's own work beginning in 1872 led to his celebrated H theorem and the Boltzmann transport equation (Fowler, 1938; Tolman, 1938). Objections about the ergodicity assumption were raised by Loschmidt (Jaynes, Levine, and Tribus, 1979; Brush, 1983), Poincaré, and Zermelo (Brush, 1983) throughout the 1890s.

To circumvent the problems of dynamics and ergodicity, J. W. Gibbs applied the method of *ensembles* to equilibrium statistical mechanics (Boltzmann, 1896; Gibbs, 1902; Fowler, 1938; Tolman, 1938). Gibbs noted that "here we may set the problem, not to follow a particular system through its succession of configurations, but to determine how the whole number of systems (an ensemble) will be distributed among the various conceivable configurations and velocities at any required time."[1] An advantage of ensemble-based reasoning was its ability to predict experimentally observed system properties without the need to invoke ergodicity. Gibbs argued that at equilibrium the classical phase-space distributions must depend only on conserved quantities, such as energy, in order to preserve the time invariance of the phase-space density (Gibbs, 1902). In addition, he noted that the phase-space density must also be non-negative and normalizable. These conditions, energy conservation and normalizability of the phase-space density, resemble Boltzmann's conditions for deriving the equilibrium distribution of particles. Gibbs noted that the exponential form of canonical equilibrium weights had "the property that when the system consists of parts with separate energies, the laws of the distribution in phase of the separate parts are of the same nature (Gibbs, 1902)." Gibbs reasoned that a closed system's phase-space distribution, while depending on all coordinates and momenta, must depend on these only through conserved quantities such as energy E from Liouville's equation. Then, by subdividing a system into two parts, having energies E_A and E_B, the phase-space distribution $\rho(A+B)$ must satisfy $\rho(A+B) = \rho(A)\rho(B)$ to be of the same "nature" for separate system parts. Gibbs showed that the only function that can satisfy this equality is the exponential $\rho(A) \propto \exp(\theta E_A)$, where both θ and the proportionality constants are independent of A. This ensemble-based logic leads to the Maxwell-Boltzmann velocity distribution (Gibbs, 1902). While Gibbs' derivation was different from Boltzmann's, its key property $\rho(A+B) = \rho(A)\rho(B)$ is the same as the key property of Boltzmann's, namely, that the multiplicity of a system is the product of multiplicities of two independent subsystems $W_{A+B} = W_A W_B$.

Presciently, Gibbs noted that (Gibbs, 1902) "although, as a matter of history, statistical mechanics owes its origin to investigations in thermodynamics, it seems eminently worthy of an independent development, both on account of the elegance and simplicity of its principles, and because it yields new results and places old truths in a new light in departments quite outside of thermodynamics." We next describe various steps taken in more recent years toward fulfilling that ambition: (1) C. Shannon (1948) developed information theory, based on a quantity that he also called entropy, $S = -\sum_i p_i \log p_i$; (2) E. T. Jaynes, assembling insights from both Gibbs and Shannon, formulated statistical physics as a tool to draw inferences about unknown quantities from limited data (Jaynes, 1957a, 1957b); (3) more recently, Shore and Johnson (1980), Livesey, Skilling and Gull, and others in the 1980s–1990s (Skilling, 1984; Skilling and Bryan, 1984; Livesey and Skilling, 1985; Livesey and Brochon, 1987; Skilling, Erickson, and Smith, 1988; Gull and Skilling, 1989;

[1] Jaynes (Jaynes, Levine, and Tribus, 1979) notes that the idea of ensembles predates Gibbs and is attributable to Maxwell.

Bryan, 1990; Skilling and Gull, 1991) broadened the intellectual landscape beyond the idea that entropy maximization was either just a matter of information or just of physics, to the perspective that entropy maximization is the only self-consistent procedure for inferring a probability distribution.

III. SHANNON'S INFORMATION THEORY AND THE MAXIMIZATION OF UNCERTAINTY

A. In information theory, entropy serves as a measure of uncertainty

In 1948, Claude Shannon considered a very different problem. Shannon was interested in the capacities of telecommunication lines to transmit information (Shannon, 1948). He wanted to minimize the average number of bits needed to encode characters in messages that were sent through noisy channels. For this problem, he derived a variational procedure that resembles the entropy maximization described above.

To illustrate Shannon's idea, consider a signal that passes through a communication channel. Suppose the message is a linear string of symbols that is M characters long, where each character is drawn independently from an r-letter alphabet, with probability p_i. Let m_i $(i = 1, 2, 3, \ldots, r)$ represent the number of times that the ith type of character is observed in the message. When M is large, the most likely M-letter message will have the composition $m_i = Mp_i$, and this occurs with probability

$$P \equiv p_1^{m_1} \cdots p_r^{m_r} = p_1^{Mp_1} \cdots p_r^{Mp_r}. \tag{6}$$

In the limiting case of an alphabet having only a single letter, the probability of knowing the message is $P = 1$, because there is only one possible string of characters. The larger the alphabet is, the smaller the value of P and the greater the uncertainty of receiving a particular message. The logarithm of $1/P$, $H = \log(1/P) = -M\sum p_i \log p_i$, is sometimes called the *uncertainty* or *missing information*. Jaynes credits Graham Wallis with similar reasoning in arguing for the mathematical form of the uncertainty (Jaynes, 2003).

Throughout this review, we use the notation $H(\{p_i\})$ not only to mean the entropy function $S(\{p_i\})$ itself. We also use this notation in derivations to denote a hypothetical functional form that will satisfy certain requirements and axioms. Shannon noted that this mathematical form of H given above is the same as the entropy function $S(\{p_i\})$ of Gibbs and Boltzmann. It can also be seen as follows. Note that there are many particular sequences consistent with the most likely message. Each of these most likely messages, having composition $m_i = Mp_i$, will occur with probability P. Other messages with $m_i \neq Mp_i$ are exponentially less likely, in the limit of large N. This result is called the asymptotic equipartition theorem (Shannon, 1948; Feinstein, 1958). If each of the most likely messages has probability P, then there are a total of $1/P$ of them. The value $1/P$ is the degeneracy W of the most likely message. Thus, maximizing W is equivalent to maximizing $1/P$. By maximizing the function H with respect to the $\{p_i\}$'s, we can predict what composition of letters $(\{p_i = p_i^*\})$ will be observed in the most probable message sent using a given alphabet. This argument shows how the maximization of the quantity $-\sum p_i \log p_i$ arises just as naturally in matters of information theory as the maximization of W arises in equilibrium physics.

Before discussing the relevance of information theory to statistical physics, we first describe the limitations of this simple argument, and efforts to find a sounder justification of $-\sum p_i \log p_i$ as a maximization principle. First, the argument above is based on the *frequentist interpretation* of probabilities. In the frequentist interpretation, a probability is estimated as the fraction of times an outcome is observed in a large number of random trials. The frequency of appearance of each face of a die is readily determined by rolling a die many times, then dividing the number of appearances of each outcome by the total number of dice rolls.

In the example above, we assumed that we see a large number of messages, from which we can compute the frequencies $p_i = m_i/M$ of the different letters. But what if only one message is seen? Such considerations are relevant to interpreting single-molecule experiments, for example, where only a single trajectory is observed.

Alternative to the frequentist interpretation is the *subjective interpretation* of probabilities. In the subjective interpretation, the concept of probability is not limited to situations that are replicable. Rather, in the subjective interpretation, a probability characterizes an observer's inference based on prior knowledge (Cox, 1961; Jaynes, 2003). In this perspective, the rules of probability are simply ways to draw inferences from premises. For instance, the probability of rain tomorrow, say p_i, is a quantity that can be estimated, even though it is not describable by a repeatable experiment. In this instance it makes no sense to speak of a ratio such as $p_i = m_i/M$ because the number of times it rains tomorrow is not an enumerable quantity. Jaynes attributes to Bernoulli (Jaynes, Levine, and Tribus, 1979) the thought that the enumeration of options "may be done in a very few cases and almost nowhere other than in games of chance the inventors of which, in order to provide equal chances to all players, took pains to set up so that the numbers of cases would be known." For the messaging problem above, we want a more general derivation that does not require observing a large number of messages.

In short, the subjective interpretation of probability is broader than the frequentist interpretation. There has been much interest in learning whether $-\sum p_i \log p_i$ and its usage as a maximization principle are also justified when probabilities are interpreted subjectively, rather than just as frequencies. Shannon developed such a derivation.

B. Shannon derived $H = -\sum p_i \log p_i$ as a measure of uncertainty

We return to the channel-capacity argument of Shannon, but now framed in the broader terms of probabilities, $\{p_i\}$, $i = 1, 2, 3, \ldots, N$, interpreted subjectively. Shannon began with some basic premises. He then derived a variational function H which, when maximized subject to certain properties of the data, returns the minimal number of bits required to represent each character. Following Jaynes (1957a), here are the three axioms that Shannon asserted must be satisfied by a proper measure of uncertainty $H(\{p_i\})$ over a set of probabilities $\{p_i\}$. The principles, described below, lead to the conclusion that H must be proportional to $-\sum_i p_i \log p_i$.

The axioms are as follows: (1) H must be a continuous function of the p_i's; (2) if the p_i's are all equal ($p_i = 1/N$), then the uncertainty H must be a monotonically increasing function of N. That is, the larger the size of the alphabet of characters, the larger the uncertainty. And (3) H must satisfy the following *composition property*:

$$H(p_1,\ldots,p_N) = H(P_1,P_2,\ldots) + P_1 H\left(\frac{p_1}{P_1},\ldots,\frac{p_{n_1}}{P_1}\right) + P_2 H\left(\frac{p_{n_1+1}}{P_2},\ldots,\frac{p_{n_1+n_2}}{P_2}\right) + \cdots, \tag{7}$$

where we regrouped terms, such that $p_1 + \cdots + p_{n_1} = P_1$, and $p_{n_1+1} + \cdots + p_{n_1+n_2} = P_2$, and so forth.

We now briefly motivate Shannon's composition property with a simple example. Following Skilling (1984), we imagine a collection of kangaroos. Kangaroos are either left handed (ℓ) or right handed (r), with probabilities p_ℓ or p_r, respectively. In addition, kangaroos are either blue eyed (b), green eyed (g), or hazel eyed (h), with probabilities p_b, p_g, or p_h, respectively. Suppose handedness is independent of eye color. We have two normalization conditions for these outcomes, $p_\ell + p_r = 1$ and $p_b + p_g + p_h = 1$. We express the uncertainty of handedness alone as the quantity $H(p_\ell, p_r)$, the uncertainty of eye color alone as $H(p_b, p_g, p_h)$, and the uncertainty of eye color and handedness combined as $H(p_{\ell b}, p_{rb}, p_{\ell g}, p_{rg}, p_{\ell h}, p_{rh})$. We refer to $p_{\ell b}, p_{rb}, p_{\ell g}, p_{rg}, p_{\ell h}, p_{rh}$ as the probabilities for *suboutcomes* to distinguish these from the probabilities for outcomes, $p_\ell, p_r, p_b, p_g, p_h$. For the function H to properly reflect our uncertainty, it must be additive across the independent variables (Livesey and Skilling, 1985),

$$H(p_{\ell b}, p_{rb}, p_{\ell g}, p_{rg}, p_{\ell h}, p_{rh}) = H(p_\ell, p_r) + H(p_g, p_b, p_h). \tag{8}$$

For instance, the uncertainty in eye color must be equal to $H(p_g, p_b, p_h)$, irrespective of whether it is applied to all kangaroos or to only left-handed kangaroos [that is, when $H(p_\ell, p_r) = 0$], since handedness and eye color are independent properties. Equation (8) expresses that eye color and handedness obey *system independence*.

Next we expand Eq. (8) using $p_\ell + p_r = 1$,

$$H(p_{\ell b}, p_{rb}, p_{\ell g}, p_{rg}, p_{\ell h}, p_{rh}) = H(p_\ell, p_r) + p_\ell H(p_g, p_b, p_h) + p_r H(p_g, p_b, p_h). \tag{9}$$

The independence of the two systems requires that $p_g = p_{\ell g}/p_\ell = p_{rg}/p_r$, $p_b = p_{\ell b}/p_\ell = p_{rb}/p_r$, $p_h = p_{\ell h}/p_\ell = p_{rh}/p_r$, leading to

$$H(p_{\ell b}, p_{rb}, p_{\ell g}, p_{rg}, p_{\ell h}, p_{rh}) = H(p_\ell, p_r) + p_\ell H\left(\frac{p_{\ell g}}{p_\ell}, \frac{p_{\ell b}}{p_\ell}, \frac{p_{\ell h}}{p_\ell}\right) + p_r H\left(\frac{p_{rg}}{p_r}, \frac{p_{rb}}{p_r}, \frac{p_{rh}}{p_r}\right). \tag{10}$$

Equation (10) says that the uncertainty over the suboutcomes $H(p_{\ell b}, p_{rb}, p_{\ell g}, p_{rg}, p_{\ell h}, p_{rh})$ is the uncertainty of the outcomes $H(p_\ell, p_r)$ plus the weighted sum of the uncertainties of the suboutcomes in each case. However, Eq. (10) is not quite Shannon's composition property. In Shannon's expression, the n_i's need not be identical; in contrast, by construction they are identical in Eq. (10). Shannon's composition property is a generalization of Eq. (10) for arbitrary regroupings of suboutcomes where suboutcomes are assumed to be independent of one another. As we will see shortly, this is related to the concept of *subset independence*.

The composition property, Eq. (7), is a recursion equation for H. We can use it to solve for H as follows. First, we define $p_i = 1/N \equiv 1/\sum_i n_i$ and $P_i = n_i/\sum_i n_i$. From Eq. (7) we have

$$H(1/N,\ldots,1/N) = H(P_1, P_2, \ldots) + \sum_i P_i H\left(\frac{1}{n_i},\ldots,\frac{1}{n_i}\right). \tag{11}$$

Defining $A(m) \equiv H(\{p_1 = 1/m, \ldots, p_m = 1/m\})$, we rewrite the above as

$$A(N) = A\left(\sum_i n_i\right) = H(P_1, P_2, \ldots) + \sum_i P_i A(n_i). \tag{12}$$

Choosing all $n_i = m$, we have

$$A(N) = A(N/m) + A(m). \tag{13}$$

Equation (13) is sufficient to specify the functional form for H. To see this, we take the derivative of Eq. (13) with respect to m to get

$$-dA(m)/dm = -\frac{N}{m^2}\frac{dA(N/m)}{d(N/m)}. \tag{14}$$

Now, substitute $m = 1$ into Eq. (14) to get

$$\frac{A'(1)}{N} = A'(N), \tag{15}$$

where $A'(1)$ is defined as $dA(m)/dm$ evaluated at $m = 1$. Solving Eq. (15) gives

$$A(N) = K \log N, \tag{16}$$

where we set the arbitrary integration constant 0 to satisfy $A(1) = 0$, and we choose $K = A'(1) > 0$ because condition 2 requires that H increase monotonically with N.

Substituting $A(N) = K \log N$ back into Eq. (12), we find

$$H(\{P_i\}) = -K\sum_i P_i \log P_i, \tag{17}$$

where $P_i = n_i/\sum_i n_i$. Shannon argues that if P_i is an irrational number, then P_i may be approximated by a rational fraction. Condition 1 (the continuity premise) thereby ensures that Eq. (17) must hold in general.

Since the variable labels $\{P_i\}$ in Eq. (17) are arbitrary, we can rewrite the above using $\{p_i\}$ instead. The state of maximum uncertainty is then predicted as the distribution that maximizes H,

$$\max H(\{p_i\}) = \max\left(-\sum_i p_i \log p_i\right), \tag{18}$$

where the maximization is with respect to each member of the set $\{p_i\}$. Equation (18) gives the same expression as the simpler frequentist argument of Boltzmann, namely, that the most probable message is simply the one that can be produced in the largest number of ways. Shannon's derivation, however, has the advantage of showing that the quantity H satisfies axiomatic properties expected of a measure of uncertainty,

even when the p_i's are not drawn from frequencies of replicable experiments and even if the p_i's are not proper probabilities [i.e., not normalized to sum to 1 (Skilling and Gull, 1991)].

IV. JAYNES REGARDED STATISTICAL PHYSICS AS A WAY TO DRAW INFERENCES FROM INCOMPLETE INFORMATION

In 1957, E. T. Jaynes brought the information-theoretic arguments of Shannon to bear on statistical physics (Jaynes, 1957a, 1957b). Shannon had already noted that (Shannon, 1948) "Quantities of the form $H = -\sum p_i \log p_i \ldots$ will be recognized as that of entropy as defined in certain formulations of statistical mechanics where p_i is the probability of a system being in cell i of its phase-space." However, Jaynes (1957a) remarked that "The mere fact that $-\sum_i p_i \log p_i$ occurs both in statistical mechanics and in information theory does not in itself establish any connection between these fields." Jaynes then argued that the business of statistical physics was to infer which particular probability distribution $\{p_i\}$ is (1) consistent with data (such as the measured average energy per particle $\bar{\varepsilon}$), and (2) that otherwise has the least possible bias with respect to all other degrees of freedom. In this way, Jaynes recast statistical mechanics as a method of inferring probability distributions from limited data.

In Jaynes' procedure, inferences are drawn by maximizing the entropy $S = -k_B \sum p_i \log p_i$ subject to constraints. Since the data are limited, many different probability distributions are consistent with the data. The choice is made by finding the set $\{p_i\}$ that both maximizes the entropy and satisfies the constraints. For instance, to infer the canonical distribution of particles, two constraints are imposed: (1) the normalization $\sum_i p_i = 1$ and (2) the given value of the average energy $\bar{E}$ estimated as $\langle E\rangle \equiv \sum_i p_i E_i$, where $\langle E\rangle$ is the theoretical expectation value based on the set $\{p_i\}$. Note that, E_i denotes energy levels of a system of particles in contrast to ε_i, used earlier, to denote single-particle energy levels. Operationally, this is done by solving the equation

$$\delta\left[S/k_B - \beta\left(\sum_i p_i E_i - \bar{E}\right) - \alpha\left(\sum_i p_i - 1\right)\right] = 0, \tag{19}$$

where the variation is with respect to each p_i and the Lagrange multipliers. As before, the Lagrange multiplier α assures the normalization of the p_i's and β enforces the known value of the average energy (which is identical to fixing the temperature in the canonical ensemble). The solution is

$$p_i^* = Q^{-1}\exp(-\beta E_i), \tag{20}$$

where $Q = \sum_i \exp(-\beta E_i)$ is the partition function for the canonical ensemble. The starred probabilities p_i^* are the values of the p_i's that maximize the entropy and satisfy the constraints. While this procedure gives exactly the same distribution law that Boltzmann obtained much earlier, Jaynes' justification for it was quite different. His derivation was based on maximizing H while satisfying data constraints, not on the basis of dynamical considerations (Jaynes, 1957a). This appears to be close to Gibbs' own earlier perspective (Gibbs, 1902). Jaynes' ideas immediately extend to the microcanonical formalism of statistical mechanics.[2]

In thermodynamics, the relevant predictor of equilibrium is not the general entropy function $S(\{p_i\})$ but rather the *post-maximization value* of the entropy

$$S_{\max} = S(\{p_i^*\}) = -k_B \sum_i p_i^* \log p_i^*$$
$$= k_B \log Q + k_B \beta \langle E\rangle = (-F + \langle E\rangle)/T,$$

where $F = -k_B T \log Q$ is the free energy, and $\beta = 1/k_B T$, obtained from $\partial S_{\max}/\partial\langle E\rangle = 1/T$.

Jaynes' information-theoretic perspective on statistical physics, however, was not unanimously embraced. Why should someone's state of knowledge or uncertainty have any bearing on physics? H, as uncertainty, was construed as a property of an observer, whereas the physical state of a system should not depend on the observer. Jaynes cites G. Uhlenbeck (Jaynes, Levine, and Tribus, 1979) who remarked that "Entropy cannot be a measure of 'amount of ignorance,' because different people have different amounts of ignorance; entropy is a definite physical quantity that can be measured in the laboratory with thermometers and calorimeters." Tikochinsky, Tishby, and Levine (1984) noted that "there are many scientists who are reluctant to use the procedure (MaxEnt) because of its reliance on the so-called subjective notion of missing information. Others consider that the concept of the entropy function should not be used outside of its original contexts."

V. SHORE AND JOHNSON: MAXIMIZING ENTROPY IS A WAY TO DRAW CONSISTENT INFERENCES

In response to these objections, another view emerged in the 1980s, due to Shore and Johnson (1980, 1981), Livesey,

[2]Using a similar reasoning to that of Eq. (19), it is also possible to fix the total energy E to infer the microcanonical distribution of statistical mechanics. In this case, the constraint is $p_i = 0$ for microstates with $E_i \neq E$

$$\sum_{i, E_i \neq E} p_i = 0, \tag{21}$$

which we rewrite as

$$\sum_i p_i(1 - \delta_{E_i,E}) = 0. \tag{22}$$

We then vary the entropy under the constraints above in addition to the normalization over the p_i's imposed using Lagrange multipliers λ and ν, respectively,

$$\delta\left\{S/k_B - \lambda\left[\sum_i p_i(1 - \delta_{E_i,E})\right] - \nu\left(\sum_i p_i - 1\right)\right\} = 0. \tag{23}$$

Variation with respect to p_i yields

$$p_i = e^{-\nu-1} \quad (E_i = E), \qquad p_i = e^{-\nu-1-\lambda} \quad (E_i \neq E). \tag{24}$$

The constraints determine the values for ν and λ. The resulting probability distribution is (Lee and Pressé, 2012b)

$$p_i = \delta_{E_i,E}/\Omega(E), \tag{25}$$

which is that expected for the microcanonical ensemble with $\Omega(E)$ denoting the total number of microstates with energy E.

Skilling, Gull, and others (Skilling, 1984; Skilling and Bryan, 1984; Livesey and Skilling, 1985; Livesey and Brochon, 1987; Skilling, Erickson, and Smith, 1988; Gull and Skilling, 1989; Skilling and Gull, 1991). Described next, this perspective embodied a transition from seeing $H = -\sum p_i \log p_i$ as a measure of uncertainty, to seeing the maximization of H as the only self-consistent way to draw inferences about probability distributions. Shore and Johnson state that "We offer an alternative approach to the problem of induction which does not involve Shannon's entropy nor any references to subjective considerations. Rather, we start from consistency conditions which must be satisfied by any algorithm for inducing a probability distribution, for a reproducible experiment. Our approach does not rely on intuitive arguments or on the properties of entropy and cross-entropy as information measures. Rather, we consider the consequences of requiring that methods of inference be self-consistent."

The approach of Shore and Johnson (1980) (SJ) represents a shift away from the view (of Shannon and Jaynes) in which H is the central object of interest to the view in which the maximum of H subject to constraints is the central object of interest.[3] SJ assert that the maximum of H with constraints should be (1) unique, (2) invariant with respect to coordinate transformation, (3) subset-independent (i.e., the relative probabilities for two subsets of outcomes within a system should not depend on other subsets if data are provided on each subset independently), and (4) system independent (i.e., the joint outcome for two independent systems must be the product of marginal probabilities if data are provided for systems independently), as described below.

The starting point for SJ is the function

$$H(\{p_i, q_i\}) - \lambda\left(\sum_i a_i p_i - \bar{a}\right), \tag{26}$$

where a is some property of interest of the distribution function p_i, $\bar{a}$ is a known average, and λ is the Lagrange multiplier that enforces the constraint.[4] The quantity q_i is the *prior distribution* for p_i (that is, the value to which each p_i defaults when no data are known) and therefore no constraints are imposed. Fixing some quantities or obtaining new knowledge can lead us to infer a distribution $\{p_i\}$ that is not equal to the prior.

We follow loosely the derivation given by SJ and others (Livesey and Skilling, 1985). Since we are only concerned with the maximum of Eq. (26) with respect to each of the $\{p_i\}$'s, we can neglect the term $\lambda\bar{a}$ above. In order to determine the functional form of H, we first invoke SJ's axiom 3, *subset independence*. Suppose that the probability of cell j is increased while that of cell k is decreased (subject to $\sum_i p_i = 1$), as a consequence of some observation or knowledge of a property of the distribution. In other words, we apply the operator $\partial_{p_j} - \partial_{p_k}$ to $H - \lambda\sum_i a_i p_i$. How is the location of the maximum of this function altered with respect to all the other cells, $l \neq k, j$? According to SJ axiom 3, this redistribution of probability among the j and k bins does not change the probability in other cells l, so

$$\partial_{p_l}(\partial_{p_j} - \partial_{p_k})\left(H - \lambda\sum_i a_i p_i\right) = 0. \tag{27}$$

Since l can label any other cell ($l \neq k, j$), Eq. (27) shows that this particular derivative $(\partial_{p_j} - \partial_{p_k})H$ must depend only on j and k. Following that logic one step further, the derivative of each such variable j and k must depend only on its own index. That is, $\partial_{p_m} H$ ($m \neq l$) must depend only on m. Thus, we find that, when H is used within a maximization procedure, it is decomposable into a sum over cell-level quantities,

$$H = \sum_i f(p_i, q_i), \tag{28}$$

where q_i is a prior for p_i which we will discuss shortly. So far, we considered a discrete system, over states i, for simplicity. However, for the next steps in the SJ argument, we switch to a continuum representation. Equation (28) becomes $H = \int \mathcal{D}[x] f(p(x), q(x))$ with $\mathcal{D}[x]$ denoting the integration measure.

Next we impose SJ's axiom 2, *coordinate invariance*. Under any coordinate transformation, $x \to y$, we have $\mathcal{D}'[y] = \mathcal{D}[x]J$, where J is the corresponding Jacobian. By assuring that both p and q remain normalized upon any coordinate transformation, we have $p' = J^{-1}p$, $q' = J^{-1}q$. In addition, in order to ensure that the constraint quantity $\int \mathcal{D}[x]p(x)a(x)$ is also preserved under coordinate transformation, we find that $a' = a$. Thus, under coordinate transformation we have $H' = \int \mathcal{D}[x]J f(J^{-1}p(x), J^{-1}q(x))$. What functional form of H leaves the maximum of $H - \lambda \int \mathcal{D}[x]p(x)a(x)$ invariant with respect to coordinate transformation? The maximum with respect to p of $H - \lambda \int \mathcal{D}[x]p(x)a(x)$ is obtained by solving

$$-\lambda a(x) + g[p(x), q(x)] = 0, \tag{29}$$

where $g[p(x), q(x)] = \partial f[p(x), q(x)]/\partial p(x)$. Next, the maximum with respect to p' of $H' - \lambda' \int \mathcal{D}'[y]p'(y)a'(y)$ is obtained by solving

$$\begin{aligned} &-\lambda' a'(y) + g[p'(y), q'(y)] \\ &\quad = -\lambda' a(x) + g[J^{-1}p(x), J^{-1}q(x)] = 0. \end{aligned} \tag{30}$$

Combining Eqs. (29) and (30) we have

$$g[J^{-1}p(x), J^{-1}q(x)] = (\lambda' - \lambda)a(x) + g[p(x), q(x)]. \tag{31}$$

The Jacobian J is an arbitrary function and λ and λ' are constants. Therefore Eq. (31) can only be true if the Jacobian vanishes from the left-hand side of Eq. (31). This happens when $g[J^{-1}p(x), J^{-1}q(x)] = g[p(x)/q(x)]$ and, therefore, $f[p(x), q(x)] = p(x)h[p(x)/q(x)] + \nu[q(x)]$, where h is thus far an unspecified function of $p(x)/q(x)$ and where we can drop the irrelevant constant ν depending only on $q(x)$.[5]

[3]We present the argument of Shore and Johnson with a sign change: their focus was a function to be minimized.

[4]Here we use a single constraint on the quantity a, and we take it to be an equality. Also, for now we consider the discrete index i, rather than the more general continuum function considered by SJ. We do this for simplicity here. These assumptions are all readily generalized. We also add that SJ label p_i their prior and q_i their posterior; this is the opposite of how these are introduced here.

[5]For a discussion on the issue of coordinate invariance, see Jaynes' discussion of transformation groups as well as discussions on Jeffreys' prior, the uninformative prior which is invariant with respect to coordinate transformation (Jeffreys, 1946, 1948; Jaynes, 1968, 2003).

Finally, we invoke SJ's axiom 4, *system independence*. Consider two independent systems described by the coordinates x_1 and x_2, with independent constraints

$$\int \mathcal{D}[x]a_k(x_k)p(x_1,x_2) = A_k \quad (k=1,2). \tag{32}$$

Defining $H = \int \mathcal{D}[x]p(x)h(r)$, where $r(x) \equiv p(x)/q(x)$ and $x \equiv \{x_1, x_2\}$, variation with respect to p of Eq. (26) gives

$$\begin{aligned}\frac{\delta}{\delta p(x)}\Big(H &- \lambda_1 \int \mathcal{D}[x]p(x_1,x_2)a_1(x_1) \\ &- \lambda_2 \int \mathcal{D}[x]p(x_1,x_2)a_2(x_2)\Big) \\ &= h[r(x)] + r(x)h'[r(x)] - \lambda_1 a_1(x_1) - \lambda_2 a_2(x_2) \\ &= h[r_1(x_1)r_2(x_2)] + r_1(x_1)r_2(x_2)h'[r_1(x_1)r_2(x_2)] \\ &\quad - \lambda_1 a_1(x_1) - \lambda_2 a_2(x_2) = 0,\end{aligned} \tag{33}$$

where $r(x)$ is equated to $r_1(x_1)r_2(x_2)$ in the last line using the system independence. Taking derivatives of the rightmost equality of Eq. (33) with respect to x_1 and x_2, yields

$$\begin{aligned}&r_1'(x_1)r_2'(x_2)[r_1^2 r_2^2 h'''(r_1 r_2) \\ &\quad + 4r_1 r_2 h''(r_1 r_2) + 2h'(r_1 r_2)] = 0,\end{aligned} \tag{34}$$

from which we obtain

$$r^2 h'''(r) + 4rh''(r) + 2h'(r) = 0. \tag{35}$$

The solution of Eq. (35) is $h(r) = -K\log(r) + B + C/r$ where K, B, C are constants. We conclude that H must have the functional form

$$H = -K \int \mathcal{D}[x]p(x)\log[p(x)/q(x)] \tag{36}$$

up to a positive multiplicative factor K, and additive constants B and C that we are free to set to zero. Alternatively, we can express this in a discrete form as $H = -K\sum_i p_i \log(p_i/q_i)$. Therefore, the SJ axioms specify that H, or any function having the same maximum as H, can be used to draw self-consistent inferences.[6] In *information geometry* the H in Eq. (36), called the *cross entropy*, is regarded as a measure of distance between two distribution functions, p and its prior q (Kullback and Leibler, 1951; Amari and Nagaoka, 2000). Maximizing H corresponds to finding the minimum distance between the two distribution functions. Others presented arguments similar to those of SJ (Livesey and Skilling, 1985; Skilling, Erickson, and Smith, 1988). Frequentist derivations allow for some simplifications; see Tikochinsky, Tishby, and Levine (1984) but see criticisms in Skilling (1984). Even the methods of SJ presented above are not entirely free of criticism. For instance, Csiszár (1991) starts from a different set of axioms in an effort to address what he believed were shortcomings of SJ, notably starting from the assumption that inference should be based on a variational principle.

In summary, on the one hand, Jaynes was criticized for justifying MaxEnt in statistical physics in terms of information, uncertainty, incomplete knowledge, and maximal ignorance. For some, those terms implied a physical reality that was not independent of the mind of an observer. Jaynes' H quantity was seen as being too grounded in Shannon's axioms about what properties would be "desirable for a measure of uncertainty." Denbigh and Denbigh (1985), for example, argued that heat capacities of materials reflect more than our lack of knowledge; heat capacities are measurable quantities.

SJ give a very different justification for entropy maximization in physics in that they regard statistical physics as an enterprise of making models for otherwise underdetermined probability distributions. This approach is different from quantifying an observer's uncertainty. In model making, we start with some initial model assumptions or priors. Those premises may be good or bad. Maximizing the entropy is seen as a procedure that enforces certain requirements for the logical consistency of nature from basic premises, not as a procedure for finding a state of maximal ignorance.

A prior distribution can reflect empirical features or intrinsic physical features of nature that are deemed relevant to the problem, such as scale invariance. For instance, for dice rolls or coin flips, an obvious choice of prior is a flat distribution. If experiments were to then show that the predictions from maximizing the entropy of a model were inadequate, it would imply the inadequacy of the prior, or that the model should be informed by additional data.

Nothing within the SJ derivation limits the applicability of entropy maximization just to systems that are at or near equilibrium. Entropy maximization is rigorous and relevant across a broad range of applications.

VI. WHAT TYPES OF CONSTRAINTS ARE APPROPRIATE FOR MaxEnt?

In applying entropy maximization principles, are there limitations on the types of constraints that can be used? We assumed that a system has some property a and that some constraints are imposed that fix the value of its average $\bar{a}$. A common textbook example is the canonical ensemble. The constraint is the temperature, which is equivalent to the average energy $\langle\varepsilon\rangle$, a first-moment quantity. Is it valid to use higher moments or combinations of moments instead? What constraints are appropriate for application within the principle of maximum entropy? Shore and Johnson express constraints in terms of the quantity $\int \mathcal{D}[x]p(x)a(x)$. We consider two aspects of this constraint quantity: how it depends on $p(x)$ and what forms of $a(x)$ are justifiable. These considerations are important for determining the full breadth of applicability of MaxEnt, including to dynamical systems.

Are constraints nonlinear in $p(x)$ appropriate for MaxEnt? SJ assert a requirement that the function H must have a

[6]Note that this derivation of $H = -K\sum_i p_i \log(p_i/q_i)$ is not in any way limited to systems that are independent of each other. Rather, it simply says that if u and v are independent of each other, then this functional form of H enforces satisfaction of the rules of addition and multiplication of probabilities for independent events, such as $p(uv) = p(u)p(v)$ for multiplication. The reason we note this is because there are instances when data are correlated. In this formalism, correlations in the model for $p(uv)$ should emerge from correlations in the data between u and v. In different formalisms, correlations were also introduced by using target functions other than $H = -K\sum_i p_i \log(p_i/q_i)$, such as the Tsallis entropy (Tsallis, 1988).

unique maximum, axiom 1. The function $\int \mathcal{D}[x]p(x)\log p(x)$ is convex, so it has a unique maximum, provided that the constraints are also convex (Shore and Johnson, 1980). The number of such constraints can be unlimited, provided these do not change the convexity of the function to be maximized and thus continue to satisfy the SJ axioms. For instance, any number of either equalities or inequalities that are linear in p_i are appropriate. In contrast, some constraints that are nonlinear in p_i may result in degenerate maxima of H, and so would violate the premises of SJ (Livesey and Skilling, 1985).

What functional forms of the property $a(x)$ are valid for MaxEnt? The axioms of SJ are found to be satisfied by many mathematical functions $a(x)$, including any polynomial of x and therefore any moment of the distribution function. However, higher moment data constraints may not provide additional independent information beyond lower moments (Shore and Johnson, 1981). Furthermore higher moments have larger associated error bars than lower moments. We discuss how we can incorporate error bars around constraints in Sec. VIII. However, for now, we point out that the larger the error bars around the data, the less this constrains the model. Next we focus on first-moment constraints, which play a particularly prominent role in physics.

A. First-moment constraints in thermodynamics are associated with large baths

Although a wide variety of constraints are used with entropy maximization principles, first-moment constraints (such as the average energy, average volume, or average particle number) often arise in equilibrium statistical physics. First-moment quantities arise because of how constraints are imposed in thermodynamic experiments. In typical experiments, a system is put into contact with a surrounding bath or reservoir, which holds the system at a fixed value of the macroscopic average, such as energy. We show below that first-moment constraints are a natural consequence of a system being in contact with an infinite bath. Higher moments used as constraints are negligible when constraints are imposed by such baths. Since the system of interest and the heat bath comprise a closed universe, the target function to be maximized can be written as

$$-\sum_{i,a} p_{ia}\log p_{ia} + \nu\left(\sum_{i,a} p_{ia} - 1\right) + \lambda\sum_{i,a} p_{ia}(1-\delta_{E_i+E_a,E_{\rm tot}}), \tag{37}$$

where indices i and a label the microstates of the open system and the heat bath, respectively, and $E_{\rm tot}$ designates the total system plus bath energy which is a fixed constant.

We now change variables so that the target function is expressed in terms of the marginal probability p_i for the open system and the conditional probability $p(a|i)$ for the heat bath, where

$$p_i \equiv \sum_a p_{ia,} \qquad p(a|i) \equiv p_{ia}/p_i. \tag{38}$$

Both probabilities satisfy the normalization conditions

$$\sum_i p_i = 1, \qquad \sum_a p(a|i) = 1. \tag{39}$$

Using Eqs. (38) and (39), the entropy in the target function becomes

$$-\sum_{i,a} p_{ia}\log p_{ia} = -\sum_i p_i\log p_i - \sum_{i,a} p(a|i)p_i\log p(a|i). \tag{40}$$

Our new target function with the entropy given by Eq. (40) constrained by Eqs. (38) and (39), with correspondingly new Lagrange multipliers, is

$$-\sum_i p_i\log p_i - \sum_{i,a} p(a|i)p_i\log p(a|i) + \sum_i \nu_i\left(\sum_a p(a|i) - 1\right) + \alpha\left(\sum_i p_i - 1\right) + \lambda\sum_{i,a} p_i p(a|i)(1-\delta_{E_i+E_a,E_{\rm tot}}). \tag{41}$$

The target function in Eq. (41) is varied with respect to $p(a|i)$, ν_i, and λ for given values of the open system variables p_i and α. From this we obtain

$$-p_i\log p(a|i) - p_i + \nu_i + \lambda p_i(1-\delta_{E_i+E_a,E}) = 0, \tag{42}$$

$$\sum_a p(a|i) = 1, \tag{43}$$

$$\sum_{i,a} p_i p(a|i)(1-\delta_{E_i+E_a,E}) = 0. \tag{44}$$

From Eq. (42) we have

$$\begin{aligned} p(a|i) &= \exp\left(\frac{\nu_i}{p_i} - 1\right) \quad (E_a = E_{\rm tot} - E_i), \\ p(a|i) &= \exp\left(\frac{\nu_i}{p_i} - 1 + \lambda\right) \quad (E_a \neq E_{\rm tot} - E_i). \end{aligned} \tag{45}$$

The constraints Eqs. (43) and (44) fix both ν_i and λ, which yields

$$p(a|i) = \frac{\delta_{E_a,E_{\rm tot}-E_i}}{\Omega(E_{\rm tot}-E_i)}, \tag{46}$$

where $\Omega(E)$ is the number of bath microstates with energy E. Substituting Eq. (46) into Eq. (41), we now have (Lee and Pressé, 2012b)

$$-\sum_i p_i\log p_i + \sum_i p_i\log\Omega_{\rm bath}(E_{\rm tot}-E_i) + \alpha\sum_i (p_i - 1). \tag{47}$$

In the limit of large bath size $E_{\rm tot} \gg E_i$, we expand $\log\Omega_{\rm bath}(E_{\rm tot}-E_i)$ to leading order in E_i. This yields $\log\Omega_{\rm bath}(E_{\rm tot}-E_i) \simeq \log\Omega_{\rm bath}(E_{\rm tot}) - \beta E_i$, where β is the inverse temperature of the bath. Therefore we see that, in this limit, the target function Eq. (47) reduces to the one given in Eq. (19), where the average energy of the system is constrained. Higher order moment constraints drop out. Exactly the same argument follows when we consider macroscopic parameters such as volume or particle numbers, which can be exchanged between the system and the environment. From this we conclude that first-order moments arise as constraints when we can control the macroscopic parameters of a system by providing contact with the environment whose size is much larger than the system under study.

Jaynes also addressed the issue of using only first-order moments by noticing equilibrium physics is usually applied to systems having large numbers of particles (Jaynes and Ford, 1963). Jaynes argued that first moments are the only quantities (per particle) that remain nonzero in the thermodynamic limit, i.e., as the system size grows large, $N \to \infty$. In particular, for a partition function $Q = \sum_E g(E)\exp(-\beta E)$, where $g(E)$ is the energy level degeneracy, the fluctuations $\langle E^2\rangle - \langle E\rangle^2$ inferred

from Q become vanishingly small with increasing system size compared with $\langle E\rangle^2$. Higher cumulants also vanish.

It follows too, however, that higher moment constraints can be useful in applications to physical systems of small size where fluctuations may be substantial (for example, in some glass-forming liquids and ferromagnetic materials) (Chamberlin, 1999, 2000; Chamberlin and Wolf, 2009; Chamberlin, Vermaas, and Wolf, 2009). "Nanothermodynamics" is a name given to the active field that studies small system sizes and finite-size effects in thermodynamics (Hill, 1962, 2001a, 2001b; Chamberlin, 2002; Balian, 2007). For the example considered above, we see that when the size of the heat bath is finite, a nonlinear constraint follows from Eq. (47),

$$\sum_i p_i f(E_i), \tag{48}$$

where $f(E) \equiv \ln\Omega_{\text{bath}}(E_{\text{tot}} - E)$ is constrained instead of the average energy. However, the general principle of entropy maximization still holds.

B. Entropies other than Gibbs-Shannon can be used when the system-independence axiom is not assumed

According to Shore and Johnson, the entropy function $-\sum_i p_i \log p_i$ satisfies the system-independence axiom. That is, the Gibbs-Shannon entropy for two independent systems (A and B) is the sum of the entropies for both systems. This gives rise to the probability multiplication rule $P(A, B) = P(A)P(B)$. However, there are systems where the assumption of system independence is inconvenient. If a system is small relative to the range of interactions in it, it may be difficult to subdivide it into independent subsystems. An example is a box that contains charged particles, where the box is smaller than the range of the interactions. These are called *nonextensive systems*. Nonextensive systems were studied in detail, particularly by Tsallis (see also Landsberg, 1972, 1984; Tsallis, 1988; Abe, 2000, 2001; Tsallis, Abe, and Okamoto, 2001; and Tsallis, Gell-Mann, and Sato, 2005). For such systems, an entropy function which is not of the form $-\sum_i p_i \log p_i$ has been suggested (Tsallis, 1988; Tsallis, Abe, and Okamoto, 2001; Tsallis, Gell-Mann, and Sato, 2005). In Shannon's terminology, nonextensive systems are characterized by an "uncertainty" quantity that is not additive,

$$H(A + B) = H(A) + H(B) + \epsilon H(A)H(B), \tag{49}$$

for subsystems A and B.

VII. ENTROPY MAXIMIZATION HAS BROAD APPLICABILITY BEYOND STATISTICAL PHYSICS

In summary, views changed over time about how to justify the principle of entropy maximization. Boltzmann recognized that the entropies that give second-law predictions of material systems at equilibrium are related to the states of maximum multiplicity of their macroscopic states. Gibbs expressed this idea in terms of ensembles of options. The arguments of both Boltzmann and Gibbs were based on the frequentist view of probabilities. Gibbs' ensembles were envisioned as imaginary replicas of all the different ways the microstates of a system could lead to the given macrostate, simply a way of satisfying a frequentist view of probabilities. Following Shannon's introduction of information theory, Jaynes reframed statistical thermodynamics as a business of making inferences about statistical systems that satisfied limited data and otherwise had the least possible bias. The foundations were reformulated yet again when Shore and Johnson proved that the principle of entropy maximization is a fundamental requirement for consistency that must be satisfied by any statistical distribution function. Maximization of entropy is now seen as broader than matters of equilibrium physics. Entropy maximization is now seen as a procedure that leads to distribution functions that satisfy basic self-consistency requirements of an inference drawn from limited data.

The broad applicability of entropy maximization can be expressed in terms of what we call *Tisza cells*. To illustrate, first consider equilibrium statistical mechanics for concreteness. Entropy maximization applies to situations involving multiple individual equivalent particles. Each particle can take on an instantaneous value of energy while the total average energy can be fixed by the temperature of a heat bath. Individual particles freely exchange energies with other particles, leading to increases or decreases in their own energies, but only in ways that leave the total energy unchanged over the whole system plus bath. Conserved properties, such as energy, can "flow" or "exchange" from one part of a system to another, since conserved properties cannot be created nor destroyed. Energies are not the only exchangeable conserved properties. Volumes and particle numbers are also conserved and exchangeable between subsystems. The power of thermodynamics comes from the ability to divide systems into subsystems, to consider the flows of conserved properties from one subsystem to another, and to predict the tendencies toward equilibrium by applying entropy maximization principles at all such scales of systems and subsystems (Tisza, 1963; Tisza and Quay, 1963; Wright, 1970). Tisza and Quay expressed this by noting that the independent variables in thermodynamics, such as particle number, volume, and energy, are "additive, conserved quantities, briefly additive invariants" (Tisza and Quay, 1963). They noted that, "Disjoint simple systems can be built up into composite systems or, conversely, we may divide a system into subsystems, sometimes referred to as cells" (Tisza and Quay, 1963). Conserved quantities can transfer, or be swapped, freely from cell to cell, subsystem to subsystem, or system to system, as when heat flows between two subsystems as the whole system approaches equilibrium. A Tisza cell can be as small as a single particle having a particular energy, to as large as a macroscopic subdivision, such as a half glass of water. Entropy maximization describes the tendencies of flows of conserved properties among Tisza cells.

Tisza cells and entropy maximization are relevant over contexts much broader than flows of energy, volume, or particles, or thermal equilibria in material systems. Not only might a Tisza cell represent one particle in a particular energy level ϵ_i, a Tisza cell could also represent a video pixel having a particular light level, an audio voxel having a particular sound level, or a dice roll having a particular score from 1 to 6.

Or a Tisza cell could represent chosen sets of pixels or dice rolls. For example, consider rolling a die many times. Suppose only the average score per roll is known. In this case, the dice-roll score is regarded as a quantity that can be swapped between one roll and any other. Maximizing the entropy S or the multiplicity W will predict the distribution of outcomes of the dice rolls. In general, any independent variable x_i is a type of "score" on one of Tisza's cells. Each cell has a numerical value x_i, for example, an energy per particle, volume per particle, etc. These score quantities x_i are not only additive, $X_N = \sum_i x_i/N$; these scores are also swappable. That is, one possible arrangement of the system, say cell i has score x_i, can be converted into another possible arrangement by "exchanging" some amount of its score with cell j. Such swaps ensure that the observable (the total value X_N) neither gains nor loses value. The quantity W then counts the numbers of different ways the total score can be distributed across the cells. MaxEnt applies to Tisza cells of any type.

VIII. MaxEnt IS USEFUL FOR MODELING IN CONJUNCTION WITH BAYES THEOREM

In this section, we describe the relationship between maximum entropy and Bayesian methods of inference (Gull and Daniell, 1978; Jaynes, Levine, and Tribus, 1979; Skilling and Gull, 1991; Caticha and Preuss, 2004). We start with Bayes' theorem, which follows from propositional logic (Cox, 1946; Aczél, 1966; Skilling and Gull, 1991; Jaynes, 2003). Given two events A and B, the probability $p(A \cap B)$ of their intersection $A \cap B$ is

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A), \tag{50}$$

where $p(A)$ is the probability of A, $p(A|B)$ is the probability of observing A given B, for example. The pair of equalities in Eq. (50) is because the set-theoretic intersection operation is symmetrical in A and B. Equation (50) is Bayes' theorem. Substituting the symbols A and B with D (for data) and M (for model) gives

$$p(M|D)p(D) = p(D|M)p(M), \tag{51}$$

where $p(M)$ is called the *model prior* and $p(D|M)$ is the probability of the data given the model, which is often called the likelihood of the model. The conditional probability $p(M|D)$ is the posterior, and $p(D)$ is the probability of the data. Equation (51) is the Bayesian framework for building mathematical models. Given data D, the goal is to search through possible models M to find the one model that maximizes the posterior, $p(M|D)$. Bayes' theorem, Eq. (51), does not tell us how to choose a functional form for any of the conditional or marginal probabilities. Bayes' theorem does not specify how such probabilities depend on data and model variables.

The role of MaxEnt here is to assert a particular choice for $p(M)$. To obtain a model probability distribution $M = \{p_i\}$ from data, MaxEnt says that the probability of a model $p(M)$ should be a monotonically increasing function of the entropy (H) of the model. That is, the greater the entropy of a model distribution function, the higher its prior probability. Since any monotonic function of the entropy will capture this property, we choose

$$p(M) \propto \exp(\alpha H), \tag{52}$$

where α is some positive constant.

The choice of $p(D|M)$ depends on the type of data available. For instance, suppose data are given as N independent quantities $\bar{a}_j$, where the index j runs from 1 to N, and each has an associated standard deviation σ_j around $\bar{a}_j$. Given a set $\{p_i\}$, the observables are estimated from theory as $\sum_i p_i a_{ij}$. The uncertainty values σ_j set an approximate bound on where the prediction $\sum_i p_i a_{ij}$ should lie (Csiszár, 1991),

$$\bar{a}_j - \sigma_j \le \sum_i p_i a_{ij} \le \bar{a}_j + \sigma_j. \tag{53}$$

As before, the ensemble average is defined as $\langle a_j \rangle \equiv \sum_i p_i a_{ij}$. This can be different than the observed average, which we denote as $\bar{a}_j$ because of the associated error bars.

Now rewrite Eq. (53) as

$$\left(\sum_i p_i a_{ij} - \bar{a}_j\right)^2 \le \sigma_j^2. \tag{54}$$

For many problems, it is reasonable to assume that errors are distributed as Gaussians. Then, the likelihood $P(D|M)$ of observing the set $\{\bar{a}_j\}$ is proportional to the product of N independent Gaussians, $\exp[-(\sum_i p_i a_{ij} - \bar{a}_j)^2/2\sigma_j^2]$, so

$$P(D|M) \propto \exp\left(-\sum_j \frac{(\sum_i p_i a_{ij} - \bar{a}_j)^2}{2\sigma_j^2}\right) \equiv \exp(-\chi^2/2). \tag{55}$$

The so-called χ^2 misfit statistic gives a smooth measure of deviation of the model prediction from the data weighted by the data uncertainty for this Gaussian model.

Fitting a model to data that has Gaussian errors, the combination of Eqs. (51), (52), and (55) gives

$$p(M|D) \propto p(D|M)p(M) \propto e^{\alpha H - (1/2)\chi^2}, \tag{56}$$

where we can drop the term $p(D)$ from the left-hand side of Eq. (56), since this term is irrelevant for the maximization with respect to the model variables M. Taking the logarithm gives

$$\log p(M|D) = -\alpha \sum_i p_i \log \frac{p_i}{q_i} + \left(-\sum_j \frac{(\sum_i p_i a_{ij} - \bar{a}_j)^2}{2\sigma_j^2}\right), \tag{57}$$

where all terms not explicitly depending on model variables were dropped. The best possible model $\{p_i\}$ is the one that maximizes Eq. (57), that is, $p(M|D)$ or $\log p(M|D)$, with respect to the variables M for a fixed value of α.

The value of α itself can be set by imposing the approximate condition

$$\sum_j \frac{(\sum_i p_i a_{ij} - \bar{a}_j)^2}{\sigma_j^2} \approx N, \tag{58}$$

which follows from a frequentist line of reasoning (Skilling, 1984).[7]

The treatment above resolves a puzzle attributed to Forney (Jaynes, Levine, and Tribus, 1979). Forney criticized MaxEnt on the grounds that it only allows the input of an average $\bar{a}_j$, whereas we should also be able to use knowledge or estimates of error bars σ_j on these averages. However, the derivation above and Eq. (57) show exactly how MaxEnt can handle error bars, if they are available, around averages. Furthermore, it is also clear from Eq. (57) that those data with larger associated error bars contribute proportionately less to the χ^2 misfit statistic. Finally, we note that although we described here only Gaussian noise, the same approach readily generalizes to other noise statistics (Meinel, 1988).

In the idealized world of perfect data where the averages are known with no error (i.e., when $\sum_i p_i a_{ij}$ is strictly equal to $\bar{a}_j$), Eq. (56) reduces to

$$\begin{aligned} p(M|D) &\propto \lim_{\{\sigma_j\}\to 0} \exp(\alpha H) \prod_j \\ &\quad \times \left[\frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(\sum_i p_i a_{ij} - \bar{a}_j)^2}{2\sigma_j^2}\right)\right] \\ &= \exp(\alpha H) \prod_j \delta\left(\sum_i p_i a_{ij} - \bar{a}_j\right), \end{aligned} \tag{59}$$

where the limit is understood to be taken after performing any integration or extremization with respect to the set $\{p_i\}$.

To see that the right-hand side of Eq. (59) is the MaxEnt recipe for constraining averages, first rewrite Eq. (59) by Fourier decomposing the Dirac delta function using the auxiliary variable k_j,

$$\begin{aligned} p(M|D) &\propto \lim_{\{\sigma_j\}\to 0} \int \prod_j dk_j \exp\Big\{\alpha H(\{p_i\}) \\ &\quad + \sum_j \left[ik_j\left(\sum_i p_i a_{ij} - \bar{a}_j\right) - \frac{\sigma_j^2 k_j^2}{2}\right]\Big\}, \end{aligned} \tag{60}$$

where we are free to bring the constant with respect to k_j, $\exp[\alpha H(\{p_i\})]$, into the integrand. The model and Lagrange multiplier which maximize the posterior $p(M|D)$ must be those which simultaneously maximize the integrand. Therefore, extremizing

$$\alpha H(\{p_i\}) + \sum_j \lambda_j \left(\sum_i p_i a_{ij} - \bar{a}_j\right) + \mathcal{N}, \tag{61}$$

with respect to both $\{p_i\}$ and $\{\lambda_j \equiv ik_j\}$, when $\sigma_j \to 0$ and where $\mathcal{N}$ is all irrelevant constants, is the MaxEnt prediction in Jaynes' limiting case that the average value is known exactly with no errors. This problem is equivalent to finding the maximum of H under the strict constraint $\sum_i p_i a_{ij} - \bar{a}_j = 0$ for each j.

[7] Skilling and Gull (1991) argued that this frequentist line of reasoning to set the constraint, Eq. (58), and thus to back out α, undermines the meticulous effort that has been put into deriving the Shannon-Jaynes measure from the self-consistent reasoning arguments of Shore and Johnson. Instead, they proposed a recipe based on a hierarchical Bayesian model to find the Lagrange multiplier that yields small improvements for the problems they examined.

A. MaxEnt is used as a tool for image reconstruction

Inferring a model by maximizing Eq. (57) has been one technique used in image reconstruction (Gull and Daniell, 1978; Skilling and Bryan, 1984; Gull and Skilling, 1989; Jaynes, 2003). In image reconstruction, the aim is to extract an image I from data D. The data and image are related by $D = F * I$, where $F * I$ denotes the transformation of the image by an operator F via some linear operation $*$. For instance, I may be an image of a moving car, F is the operator that describes how the camera blurs that image, and D is the blurry image of the moving car (Skilling and Gull, 1991; Steinbach *et al.*, 1992). In principle, obtaining the image I from the data D should be simple. If the transform is invertible, we have $I = (F*)^{-1}D$. In reality because the data are noisy, the operation is ill-conditioned. MaxEnt provides one way of regularizing this inversion.

As an example, MaxEnt has been used to obtain real-space images from x-ray scattering data (Gull and Daniell, 1978). The data are in the form of a probability density in Fourier space $\{p_k\}$, the image is in real space $\{p_x\}$, and the convolution operator is the Fourier transform $\mathcal{F} * I \equiv \sum_x \exp(i2\pi kx/N) * I$. Here the analog of Eq. (57) is (Gull and Daniell, 1978)

$$\log p(M|D) = -\alpha \sum_x p_x \log \frac{p_x}{q_x} + \left(-\sum_k \frac{|\mathcal{F} p_x - p_k|^2}{2\sigma_k^2}\right). \tag{62}$$

Other regularization methods also exist, such as the commonly used method of Tikhonov regularization (Engl, Kunisch, and Neubauer, 1989), although such methods are not consistent with the axioms of Shore and Johnson.

Image-reconstruction methods were used to extract models from different types of spectroscopic data. For instance, consider a noisy fluorescence decay signal $f(t)$. It is possible to express $f(t)$ as follows:

$$f(t) = \int_0^\infty d\tau \alpha(\tau) e^{-t/\tau}, \tag{63}$$

where $\alpha(\tau)$ is a distribution of decay rates, interpretable as the image I. In this case, the linear operator that relates $\alpha(\tau)$ to $f(t)$ is the Laplace operator. The signal decay of the experiment $I(t)$ is given by

$$I(t) = E(t) * f(t) = E(t) * \int_0^\infty d\tau \alpha(\tau) e^{-t/\tau}, \tag{64}$$

where $E(t)$ is the temporal shape of the excitation pulse relating the observed decay intensity to the fluorescence decay curve.

The MaxEnt approach infers $\alpha(\tau)$ from a noisy decay curve. The advantage of MaxEnt here is that it does not impose features on this distribution of rates that are not otherwise warranted by the data (Livesey and Brochon, 1987). It does not require prior knowledge of how many exponential components contribute to the decay. Using Eq. (57), we have

$$\log p(M|D) = -\int d\tau \alpha(\tau) \log\left(\frac{\alpha(\tau)}{m(\tau)}\right) - \lambda \int dt \left(\frac{[I_{\text{theo}}(t) - I_{\text{obs}}(t)]^2}{2\sigma(t)^2}\right), \quad (65)$$

where $I_{\text{theo}}(t)$ is the theoretical intensity which is set equal to $E(t) * \int_0^\infty d\tau \alpha(\tau) e^{-t/\tau}$ and $I_{\text{obs}}(t)$ is the observed intensity, λ is the Lagrange multiplier, and $m(\tau) = 1/\tau$ is the scale-invariant prior.

Livesey and Brochon (1987) probed the fluorescence of L-tryptophan at two emission wavelengths, 390 and 320 nm. The unexpected outcome, shown in Fig. 1, was that $\alpha(\tau)$ is qualitatively different at different wavelengths. The power of the method of image reconstruction is that the decay curve need not be fit to one, two, three, or more exponentials. Rather, the data are inverted directly, minimizing the biases that would otherwise be introduced in more traditional modeling.

MaxEnt has been widely useful in physical modeling: finding the continuous distribution of rates of rebinding a carbon monoxide ligand to a heme protein (Steinbach *et al.*, 1992); measuring the static and dynamic properties of binding flavin adenine dinucleotide (FAD) ligand to flavin reductase protein (Yang *et al.*, 2003); measuring the complex folding kinetics of dihydrofolate reductase (Steinbach, Ionescu, and Matthews, 2002); determining the fluorescence of donor-acceptor distance distribution function for the conformational states of a simple polymer, poly-(L-proline), from single-molecule Foerster resonance energy transfer (FRET) data (Watkins, Chang, and Yang, 2006); inferring the firing patterns of arrays of neuronal cells (see below) (Schneidman *et al.*, 2006); inferring the structures of networks of proteins (Locasale and Wolf-Yadlin, 2009) and genes (Lezon *et al.*, 2006) from proteomics and microarray data; modeling low-noise regulatory networks assuming there is maximal information transmission from transcription factors (input) to gene products (output) (Tkacik, Walczak, and Bialek, 2009; Walczak, Tkacik, and Bialek, 2010); inferring diffusion coefficient distributions from fluorescence correlation spectroscopy data (Sengupta *et al.*, 2003); and modeling the apparent search strategy, called "infotaxis," of moths seeking the source of pheromones that are delivered in bursts, rather than from a stable gradient (Vergassola, Villermaux, and Shraiman, 2007).

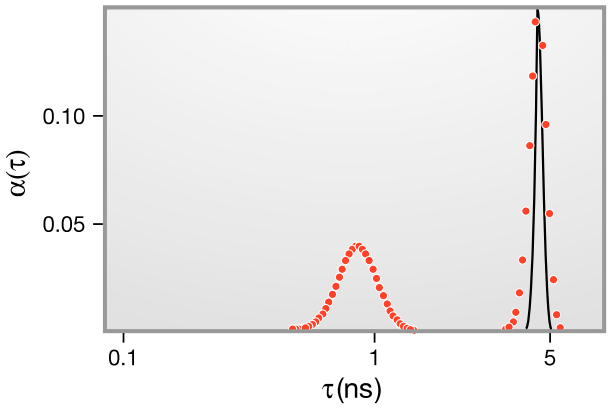

FIG. 1 (color online). The rate distribution $\alpha(\tau)$ taken from pulse fluorimetry experiments on L-tryptophan - vs τ on a logarithmic scale at two emission wavelengths (390 nm solid line, 320 nm with circles). From Livesey and Brochon, 1987.

IX. MAXIMUM CALIBER IS THE MAXIMUM-ENTROPY PRINCIPLE APPLIED TO DYNAMICAL PATHWAYS

The developments of Shore and Johnson show that the principle of entropy maximization is not limited to material particles or states of equilibrium. Entropy maximization is as rigorously applicable to computing the probabilities of dynamical pathways, often called the principle of maximum caliber (MaxCal) (Jaynes and Haken, 1985).

Rather than seeking distributions of equilibrium states, we seek probability distributions over dynamical trajectories. Before giving the details, here is an overview. In equilibrium physics, many possible probability distributions $p(E)$ over energies E are consistent with an observed average energy. The preferred probability distribution $p(E)$ is that which is inferred by maximizing an entropy over microstates, subject to a known value of an observable, such as the average energy. Now, for nonequilibrium physics, the maximum caliber approach infers the probabilities of different possible trajectories by maximizing a "route entropy" over all the possible dynamical pathways, subject to a known value of an observable dynamical quantity, say an average velocity or flux. To give some intuition, consider all the roads from New York to Chicago as a metaphor for all the possible dynamical paths from one physical state to another. Suppose a single quantity is known, namely, the average rate at which cars reach Chicago from New York. The problem is then to predict the distribution of fluxes of cars through all the possible routes.

For dynamical problems of this type, we define the *path entropy* as

$$H(\{p_C\}) = -\sum_C p_C \log p_C, \quad (66)$$

where p_C is the probability that the dynamical process follows one particular path C. We suppose that there are constraints on the dynamics, indexed by α,

$$F^{(\alpha)}(p_C) = 0, \quad (67)$$

where α runs from 1 to the total number of constraints. While this formalism is valid for general forms of the constraint, Eq. (67), following Shore and Johnson (1980), we are often interested in constraints that are linear in p_C (such as average fluxes, velocities, or rates of conversion),

$$\sum_C A_C^{(\alpha)} p_C - \bar{A}^{(\alpha)} = 0, \quad (68)$$

where $\bar{A}^{(\alpha)}$ is the measured average of the quantity $A_C^{(\alpha)}$ over the paths C.

In order that p_C quantities represent proper probabilities, one constraint that must be satisfied is the normalization over the path probabilities,

$$\sum_C p_C - 1 = 0. \quad (69)$$

The principle of MaxCal is to maximize the entropy over pathways, Eq. (66), subject to constraints given by Eq. (68), yielding the probability distribution over pathways,

$$p_C = \frac{\exp(\sum_\alpha \lambda_\alpha A_C^{(\alpha)})}{Q}, \tag{70}$$

where the $\{\lambda_\alpha\}$ are the Lagrange multipliers that enforce the constraints and

$$Q(\{\lambda\}) = \sum_C \exp\left(\sum_\alpha \lambda_\alpha A_C^{(\alpha)}\right) \tag{71}$$

is the sum of statistical weights over pathways, called the *dynamical partition function*. The dynamical partition function plays the same role here that the equilibrium partition function plays in equilibrium statistical mechanics. MaxCal is a procedure for predicting the relative probability that a system will take trajectory C from one physical state to another. It is not approximate, and is not limited to near-equilibrium processes. It follows from the derivation of Shore and Johnson that MaxCal is as sound a basis for nonequilibrium processes as the maximization of entropy is for equilibrium.[8]

Equations (70) and (71) are formalistic. They are not explicitly computable unless the set of pathways is specified. Jaynes was interested in continuous pathways C that satisfy deterministic Hamiltonian equations of motion. However, for continua, constructing a path ensemble using the microscopic dynamics is difficult. Although some formal relations in linear nonequilibrium thermodynamics (Onsager and Machlup, 1953a,1953b) were derived from such a formalism (Jaynes, Levine, and Tribus, 1979; Jaynes and Haken, 1985; Haken, 1986; Dewar, 2005, 2009), the practical applicability has been limited. As Jaynes (Jaynes and Haken, 1985) put it, "It is probably beyond our mathematical ability to do the indicated calculations explicitly for any really nontrivial problem; that is perhaps a task for the computers of the next century." However, MaxCal is readily applied to systems having discrete dynamical states, as we show below.

A. Filyukov and Karpov introduced the maximization of path entropies over discrete paths

A practical approach to path entropy maximization was formulated early by Filyukov and Karpov (1967a, 1967b) and Filyukov (1968). Closely related ideas were independently developed by Zubarev (Zubarev and Zubarev, 1961; Zubarev, 1971), Attard (2009), and later proposed by Evans and coworkers (Evans, 2004a, 2004b, 2005). Rather than continuous paths, Filyukov and Karpov considered trajectories composed of discrete time steps. They also considered coarse-grained trajectories that follow a stochastic dynamics, instead of following deterministic Hamiltonian equations of motion. Such methods also were developed by others (Gaspard, 2004; Lecomte, Appert-Rolland, and van Wijland, 2007; Monthus, 2011; Smith, 2011). Following Jaynes (Jaynes and Haken, 1985), we collectively refer to these path entropy maximization methods as MaxCal.

In this approach, the observable dynamical properties of a system are estimated from MaxCal as averages over different discrete pathways the system can take. We call the collection of these pathways an ensemble of paths. A particular pathway having a total of T time steps is described by a sequence $C = \{i_0, i_1, \ldots i_T\}$, where i_x is the state occupied by the system at time x (Filyukov and Karpov, 1967a,1967b; Filyukov, 1968). An example of such a pathway is shown in Fig. 2.

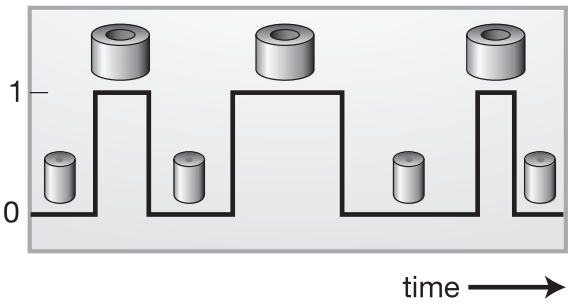

FIG. 2. One possible stochastic trajectory of a single ion-channel transitioning between open (conducting) and closed (nonconducting) states (Phillips, Kondev, and Theriot, 2009; Dill and Bromberg, 2011).

The path entropy for a discrete trajectory having probability $p_C = p_{i_0,\ldots,i_T}$ is given by Eq. (66). We refer to a particular sequence of states visited in time $i_0, \ldots, i_T$ as a microtrajectory. For the example given in Fig. 2, the sum in the path entropy formula is over all paths which coincide with all states (open or closed, in this case) that can be occupied at the discrete times $0, 1, \ldots, T$. The entropy for the discrete time process, the discrete time analog of Eq. (66), is then

$$H(T) = -\sum_{i_0,i_1,\ldots,i_T} p_{i_0 i_1,\ldots,i_T} \log p_{i_0 i_1,\ldots,i_T}. \tag{72}$$

Filyukov and Karpov (1967a, 1967b) and Filyukov, (1968) assumed a stationary first-order Markov process,

$$p_C = p_{i_0} p_{i_0 \to i_1} p_{i_1 \to i_2} \cdots p_{i_{T-1} \to i_T}, \tag{73}$$

where $p_{i\to j}$ is a transition probability from state i to state j and p_i is a single state occupation probability. Inserting Eq. (73) into Eq. (72) gives an expression for the path entropy,

$$H(T) = -\sum_{\{i_0,i_1,\ldots i_T\}} p_{i_0} p_{i_0 \to i_1} p_{i_1 \to i_2} \cdots p_{i_{T-1} \to i_T} \times \log p_{i_0} p_{i_0 \to i_1} p_{i_1 \to i_2} \cdots p_{i_{T-1} \to i_T}. \tag{74}$$

Expressing the logarithm of the product in Eq. (74) as a sum over logarithmic terms we obtain

$$H(T) = -\sum_{\{i_0,i_1,\ldots i_T\}} p_{i_0} p_{i_0 \to i_1} p_{i_1 \to i_2} \cdots p_{i_{T-1} \to i_T} \log p_{i_0} - \sum_{\{i_0,i_1,\ldots i_T\}} p_{i_0} p_{i_0 \to i_1} p_{i_1 \to i_2} \cdots p_{i_{T-1} \to i_T} \times \log p_{i_0 \to i_1} + \cdots. \tag{75}$$

We now simplify the entropy above for the special case of stationary state occupation and transition probabilities.[9] We do so by performing multiple sums in Eq. (75) using the following normalization and balance conditions:

$$\sum_j p_{i\to j} = 1, \qquad \sum_i p_i p_{i\to j} = p_j. \tag{76}$$

[8]The term MaxCal was coined by Jaynes (Jaynes and Rabinovitch, 1980; Jaynes and Haken, 1985), with a meaning that is loosely related to the bore diameters of guns, from which the term caliber derives.

[9]See Lee and Pressé (2012a) for a derivation for when state occupation and transition probabilities can be assumed stationary.

In the first term on the right-hand side of Eq. (75), the summation over all indices except i_0 yields $-\sum_i p_i \log p_i$ where we simplified the notation on the dummy indices. In the second term, all indices except i_0 and i_1 can be summed over. The second term reduces to $-\sum_{i,j} p_i p_{i\to j} \log p_{i\to j}$. All other terms on the right-hand side of Eq. (75) are treated similarly to the second.

This yields

$$H(T) = -\sum_i p_i \log p_i - T\sum_{i,j} p_i p_{i\to j} \log p_{i\to j}. \tag{77}$$

For large T, the first term is negligible and the path entropy is approximated as

$$H(T) = TH_1, \tag{78}$$

where

$$H_1 = -\sum_{i,j} p_i p_{i\to j} \log p_{i\to j} \tag{79}$$

is the path entropy per step.

In contrast, when constraints are imposed on singlet-state occupation probabilities, as is the case of equilibrium statistical mechanics, the entropy is

$$H(T) = -T\sum_i p_i \log p_i. \tag{80}$$

Comparing Eqs. (80) and (79), we find that the entropy of Eq. (79) reduces to Eq. (80), when

$$p_{i\to j} = p_j. \tag{81}$$

Physically, Eq. (81) is interpreted as instant equilibration.

Next, we consider the more interesting situation, a reversal of this logic. That is, rather than assuming Markov-chain kinetics and deriving the form of path entropy, we begin by assuming MaxCal (the maximization of path entropy) to be the most basic prediction principle of dynamics, and we show that the nature of what experimental data are observed dictates whether a process can be modeled as Markovian or not. If data take the form of two-point constraints (for example, the numbers of point-to-point transition events), then it follows that first-order Markov models are the unique solution to the question of what type of model maximizes the caliber. It says that Markov models are not the first principle; MaxCal and the form of the experimental data are the first principle. It says that the structure of the data is sufficient to dictate how the dynamical process should be modeled. Going further, if the data appear instead in the form of three-point information or higher, then maximizing the caliber uniquely infers increasingly complex models. Thus MaxCal gives a principled and systematic way of modeling dynamics directly derived from the underlying nature of the data themselves (Ge *et al.*, 2012; Lee and Pressé, 2012a). This derivation is given below.

B. Markov processes follow from the principle of MaxCal

What is the justification for asserting a Markov model as the first step in modeling a kinetic process? We show here that first-order Markov processes are the unique solution to the question of what types of physical processes maximize the caliber (i.e., that maximize the path entropy subject to constraints) when the data count the number of transitions (Ge *et al.*, 2012); see Lee and Pressé (2012a) for a more general discussion on pairwise constraints and Markovianity.

For a discrete time process, the path entropy is

$$H(T) = -\sum_{i_0,i_1,\ldots,i_T} p_{i_0 i_1,\ldots,i_T} \log p_{i_0 i_1,\ldots,i_T}. \tag{82}$$

Now, we impose pairwise constraints for each step $m \to n$ over the time period $[0, T]$, i.e.,

$$\langle N_{m\to n}\rangle = \sum_{i_0,\ldots,i_T} p_{i_0\ldots,i_T} N_{m\to n}(i_0,\ldots,i_T), \tag{83}$$

where $N_{m\to n}(i_0,\ldots,i_T) \equiv \sum_{k=0}^{T-1} \delta_{i_k,m}\delta_{i_{k+1},n}$ counts the number of $m \to n$ transitions. We verify that $\sum_{m,n} N_{m\to n} = T$.

We then maximize the path entropy, Eq. (72), with the constraints given by Eq. (83) using λ_{mn} as the Lagrange multiplier to constrain $\langle N_{m\to n}\rangle$. This yields

$$p_{i_0,\ldots,i_T} = \prod_{k=0}^{T-1} p_{i_k\to i_{k+1}} \propto e^{-\sum_{m,n}\lambda_{mn}\sum_{k=0}^{T-1}\delta_{i_k,m}\delta_{i_{k+1},n}}, \tag{84}$$

where, from the second proportionality, we have $p_{i_k\to i_{k+1}} \propto e^{-\lambda_{i_k i_{k+1}}}$ and the probability $p_{i_k\to i_{k+1}}$ is understood as the conditional probability $p(i_{k+1}|i_k)$. Thus, under the constraints imposed by Eq. (83), the joint probability distribution $p_{i_0\cdots i_T}$ given by Eq. (84) is a first-order Markov process. That is, it can be rewritten as the product of transition probabilities which describe the probability of being in a state at some time $k+1$ as depending only on the state at time k.

The following dynamical partition function determines the proportionality constant of Eq. (84):

$$Q(T) = \sum_{i_0,\ldots,i_T} e^{-\sum_{m,n}\lambda_{mn}\sum_{k=0}^{T-1}\delta_{i_k,m}\delta_{i_{k+1},n}}. \tag{85}$$

Derivatives of this dynamical partition function readily yield quantities such as the average number of $m \to n$ transitions, $\langle N_{m\to n}(i_0,\ldots,i_T)\rangle = -\partial \log Q(T)/\partial\lambda_{mn}$. This resembles the way that taking derivatives of equilibrium partition functions yield equilibrium averages as well as higher cumulants.

C. MaxCal resembles MaxEnt in its mathematical structure: partition functions and their derivatives

The procedural logic of MaxCal for path distributions is similar to that of MaxEnt. MaxEnt starts with experimental constraints. Maximizing the entropy over microstates gives the Boltzmann statistical weights for the accessible states parametrized by data. The partition function is the sum over these statistical weights. Derivatives of the partition function give information about moments of the distribution that were not used to parametrize the model.

Similarly MaxCal starts with experimental constraints. Maximizing the entropy over pathways gives statistical weights for the various trajectories. The dynamical partition function is the sum over these statistical weights. Derivatives of this partition function give information about moments of the dynamical distribution. Next, we show this in more detail. To keep it simple, we derive results only for Markov processes.

We can generalize our previous discussion on pairwise statistics to include singlet constraints as well, which we

constrain using Lagrange multipliers $\{\alpha\}$. For the trajectory probabilities, this yields

$$p_{i_0 i_1 \ldots i_T} \propto e^{-\sum_m \alpha_m \sum_{k=0}^{T} \delta_{i_k,m} - \sum_{m,n} \lambda_{mn} \sum_{k=0}^{T-1} \delta_{i_k,m} \delta_{i_{k+1},n}}. \tag{86}$$

Singlet and pairwise constraints have been interpreted as constraints on energy and flux, respectively, in the literature (Monthus, 2011). For systems having more complex behavior, for which singlet or pairwise statistics may be correlated, such correlations can also be included as constraints for modeling the dynamics (Schneidman *et al.*, 2006; Pressé, Ghosh, and Dill, 2011).

The dynamical partition function for Eq. (86) is the sum over path weights,

$$\begin{aligned} Q(\{\alpha_m, \lambda_{mn}\}) &= \sum_{i_0 i_1 \ldots i_T} e^{-\sum_m \alpha_m \sum_{k=0}^{T} \delta_{i_k,m} - \sum_{m,n} \lambda_{mn} \sum_{k=0}^{T-1} \delta_{i_k,m} \delta_{i_{k+1},n}} \\ &= \sum_{i_0 i_1 \cdots i_T} e^{-\sum_m \alpha_m N_m(i_0,\ldots,i_T) - \sum_{m,n} \lambda_{mn} N_{m\to n}(i_0,\ldots,i_T)}, \end{aligned} \tag{87}$$

where $\sum_{k=0}^{T} \delta_{i_k,m}$ is, as before, the total count of dwells in state m over a total time T, which we denote $N_m(i_0, \ldots, i_T)$. Likewise, $\sum_{k=0}^{T-1} \delta_{i_k,m} \delta_{i_{k+1},n}$ is the total number of transitions from states m to n over a total time T which we denote $N_{m\to n}(i_0, \ldots, i_T)$.

The dynamical partition function above can be rewritten using standard transfer matrix methods (Monthus, 2011),

$$Q(\{\alpha_m, \lambda_{mn}\}) = \mathbf{v}^\dagger \cdot \mathbf{G}^T \cdot \mathbf{v}, \tag{88}$$

where

$$v_i = \exp(-\tfrac{1}{2}\alpha_i), \qquad G_{ij} = \exp[-\tfrac{1}{2}(\alpha_i + \alpha_j) - \lambda_{ij}]. \tag{89}$$

$\mathbf{G}^T$ denotes the transfer matrix raised to the Tth power and $\mathbf{v}^\dagger$ denotes the transpose of $\mathbf{v}$.

Initial and final conditions are unspecified in the dynamical partition function, Eq. (88). We can specify these as additional constraints just as we constrained observed transitions (Lee and Pressé, 2012a). In the dynamical partition function, Eq. (88), specifying an initial condition is equivalent to replacing $\mathbf{v}^\dagger$ by a specified row vector $\mathbf{a}^\dagger$ as follows:

$$Q(\{\alpha_m, \lambda_{mn}\}) = \mathbf{a}^\dagger \cdot \mathbf{G}^T \cdot \mathbf{v}. \tag{90}$$

Similarly, arbitrary final conditions can be incorporated by replacing $\mathbf{v}$ by an arbitrary column vector $\mathbf{b}$.

Fluctuations as well as higher order cumulants of relevant physical quantities such as $N_m(i_0, i_1, \ldots, i_T)$ and $N_{m\to n}(i_0, i_1, \ldots, i_T)$ are then readily inferred as higher derivatives of the dynamical partition function,

$$\langle N_m(i_0, \ldots, i_T)^k \rangle_c = (-)^k \frac{\partial^k}{\partial \alpha_m^k} \log Q(\{\alpha_m, \lambda_{mn}\}), \tag{91}$$

$$\langle N_{m\to n}(i_0, \ldots, i_T)^k \rangle_c = (-)^k \frac{\partial^k}{\partial \lambda_{mn}^k} \log Q(\{\alpha_m, \lambda_{mn}\}), \tag{92}$$

where the subscripted c is used to denote cumulants.

Following reasoning similar to that of Sec. VI, in the limit of long trajectories $T \to \infty$, second and higher moment constraints will not substantially contribute to determining the model for the trajectory distribution. Hence, dynamical constraints in MaxCal will take the form of simple first-moment averages, such as the average flux $\langle J \rangle$, provided the trajectories are long enough.

We return to the metaphor of Tisza cells. We noted earlier that entropy and its maximization apply to any type of Tisza cell: dice rolls having scores, pixels having light intensities, or messages composed from alphabets, just as readily as it applies to particles having equilibrium energies or volumes. Here we can consider our Tisza cell a single time step in the trajectory of a particle. The maximization of the path entropy in the MaxCal procedure simply ensures that the predicted pathway probability distribution factorizes into independent probabilities when a constraint does not couple paths. However, if time steps are much shorter than typical relaxation times of the particles, the data in different Tisza cells can be highly correlated over distant parts of the trajectory. In thermodynamics, it is in these types of situations where entropies such as Tsallis' entropy which do not satisfy the system-independence axiom (Tsallis, 1988; Tsallis, Abe, and Okamoto, 2001; Tsallis, Gell-Mann, and Sato, 2005) were applied. Furthermore time correlations or spatial correlations, from which memory in the system emerges, can also be used as constraints and are important aspects of the range of applicability of MaxCal (see the toggle switch in Sec. XI.E, for example); see also Harris and Touchette (2009) and Ge *et al.* (2012).

D. The master equation follows from the principle of MaxCal

A common strategy in modeling kinetic processes is to begin by asserting a master equation, then computing resulting dynamical properties of interest. Here we follow a logic similar to Sec. IX.B on Markov processes and show that the master equation follows from MaxCal.

As discussed in Sec. IX.B, MaxCal under pairwise constraints results in a first-order Markov process. However, the resulting transition probability $p(a \to b; t)$ and the state occupation probability $p(a; t)$ predicted from MaxCal are time dependent in general (Lee and Pressé, 2012a). The master equation describes the situation where the transition probability $p(a \to b)$ is time independent, but the occupation probability $p(a; t)$ is time dependent. We derive the master equation by generalizing the previous arguments and keeping all time dependence explicit.

We start with the joint probability distribution $p_{i_0 i_1, \ldots, i_T}$. Under singlet and pairwise constraints, the joint probability distribution is expressed using transfer matrix notation introduced in Sec. IX.C,

$$p_{i_0 i_1, \ldots, i_T} = \frac{v(i_0) G(i_0, i_1) G(i_1, i_2) \cdots G(i_{T-1}, i_T) v(i_T)}{\mathbf{v}^\dagger \cdot \mathbf{G}^T \cdot \mathbf{v}}. \tag{93}$$

The m-point joint probability distribution is obtained from Eq. (93) by summing over indices $i_0, \ldots, i_{t-m}, i_{t+1}, \ldots, i_T$ as follows:

$$p(a_1,\ldots,a_m;t) \equiv \sum_{i_0,\ldots,i_{t-m},i_{t+1},\ldots,i_T} p(i_0,i_1,\ldots,i_{t-m},a_1,\ldots,a_m,i_{t+1},\ldots,i_T)$$
$$= \frac{[\mathbf{v}^\dagger \mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)G(a_2,a_3)\cdots G(a_{m-1},a_m)[\mathbf{G}^{T-t}\mathbf{v}](a_m)}{\mathbf{v}^\dagger \mathbf{G}^T \mathbf{v}}$$
$$= \frac{[\mathbf{v}^\dagger \mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)G(a_2,a_3)\cdots G(a_{m-1},a_m)[\mathbf{G}^{T-t}\mathbf{v}](a_m)}{\mathbf{v}^\dagger \mathbf{G}^T \mathbf{v}}, \tag{94}$$

where $[\mathbf{v}^\dagger \mathbf{G}^n](a)$ and $[\mathbf{G}^n\mathbf{v}](a)$ denote the ath components of the row and column vectors $\mathbf{v}^\dagger\mathbf{G}^n$ and $\mathbf{G}^n\mathbf{v}$, respectively. Similarly we denote $[\mathbf{G}^n](a,b)$ the (a,b) component of the matrix $\mathbf{G}^n$. For notational convenience, since many of the quantities considered here are explicitly time dependent, we bring the state labels from the subscript into the main brackets, namely, we write $p(a_1,\ldots,a_m;t)$ not $p_{a_1,\ldots,a_m}(t)$.

The time index in $p(a_1,\ldots,a_m;t)$ is required, as this probability depends on which indices in Eq. (94) are summed over. As before, we obtain conditional or transition probabilities from joint probabilities as follows:

$$p(a_1,\ldots,a_m;t)p(a_1,\ldots,a_m \to a_{m+1};t)$$
$$= p(a_1,\ldots,a_{m+1};t+1). \tag{95}$$

Combining Eqs. (94) and (95), we have

$$p(a_1,\ldots,a_m \to a_{m+1};t) = \frac{[\mathbf{v}^\dagger \mathbf{G}^{t-m}](a_1)G(a_1,a_2)\cdots G(a_m,a_{m+1})[\mathbf{G}^{T-t}\mathbf{v}](a_{m+1})}{[\mathbf{v}^\dagger \mathbf{G}^{t-m}](a_1)G(a_1,a_2)\cdots G(a_{m-1},a_m)[\mathbf{G}^{T-t+1}\mathbf{v}](a_m)}$$
$$= \frac{G(a_m,a_{m+1})[\mathbf{G}^{T-t}\mathbf{v}](a_{m+1})}{[\mathbf{G}^{T-t+1}\mathbf{v}](a_m)}$$
$$= p(a_m \to a_{m+1};t). \tag{96}$$

Thus, under singlet and pairwise constraints, the transition probability $p(a_1,\ldots,a_m \to a_{m+1};t)$ given by Eq. (96) reduces to that expected for a first-order Markov process, $p(a_m \to a_{m+1};t)$. We made explicit above the time dependence of the transition probability.

To derive the master equation, we must know under what conditions state occupation probabilities are time dependent and transition probabilities are time independent. To answer this question we apply the Perron-Frobenius theorem to the $\mathbf{G}$ transfer matrix, a square matrix of size $N \times N$ with positive elements. According to the theorem, $\mathbf{G}$ satisfies the following properties:

(1) $\mathbf{G}$ has a positive real eigenvalue r such that any other eigenvalue λ is strictly smaller than r in absolute value.
(2) There is a left eigenvector $\mathbf{y}$, $\mathbf{y}^\dagger\mathbf{G} = r\mathbf{y}^\dagger$, where $y_i > 0$ for all i. There is also a corresponding right eigenvector $\mathbf{z}$, where $\mathbf{G}\mathbf{z} = r\mathbf{z}$ and $z_i > 0$ for all i.
(3) Left and right eigenvectors with eigenvalue r are nondegenerate.
(4) $\lim_{T\to\infty}(\mathbf{G}^T/r^T) = \mathbf{z}\mathbf{y}^\dagger$.

The vector $\mathbf{v}$ has only nonnegative elements. From point (4) above we have

$$\lim_{T\to\infty}\frac{\mathbf{G}^T\mathbf{v}}{r^T} = \mathbf{z}(\mathbf{y}^\dagger\mathbf{v}); \qquad \lim_{T\to\infty}\frac{\mathbf{v}^\dagger\mathbf{G}^T}{r^T} = (\mathbf{v}^\dagger\mathbf{z})\mathbf{y}^\dagger. \tag{97}$$

Inserting Eq. (97) into Eq. (96) in the limit that $T - t \to \infty$, we recover

$$p(a \to b) = \frac{G(a,b)z(b)}{rz(a)}. \tag{98}$$

That is, the transition probability is time independent in this limit. However, from Eq. (94), the m-point joint probabilities remain time dependent when $T - t$ is large,

$$p(a_1,\ldots,a_m;t) = \frac{[\mathbf{v}^\dagger \mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)G(a_2,a_3)\cdots G(a_{m-1},a_m)z(a_m)}{r^t\mathbf{v}^\dagger\mathbf{z}}, \tag{99}$$

and, in particular, this is true for the one-point occupation probability

$$p(a;t) = \frac{[\mathbf{v}^\dagger \mathbf{G}^t](a)z(a)}{r^t\mathbf{v}^\dagger\mathbf{z}}. \tag{100}$$

Thus we can maximize the path entropy under singlet and pairwise constraints for a trajectory of infinite length. For a finite time duration, when $(T - t \to \infty)$, we obtain a time-homogeneous Markov process which is here described by (1) time-independent transition probabilities and (2) time-dependent one-point occupation probabilities.

From Eqs. (98) and (100), we arrive at the evolution equation for the time-dependent one-point occupation probability evolved according to time-independent transition probabilities,

$$p(a;t+1) = \sum_b p(b;t)p(b \to a). \tag{101}$$

Subtracting $p(a;t)$ from both sides of Eq. (101), we get

$$p(a;t+1)-p(a;t)=\sum_b p(b;t)p(b\to a)-p(a;t). \quad (102)$$

Substituting into Eq. (102), the normalization condition given by Eq. (76) yields

$$p(a;t+1)-p(a;t)=\sum_b p(b;t)p(b\to a) - \sum_b p(a;t)p(a\to b). \quad (103)$$

Taking the continuum limit by replacing $t+1$ for $t+\Delta t$ in Eq. (103) and defining the transition rates to be $k_{b\to a}\equiv p(b\to a)/\Delta t$, we get

$$\frac{p(a;t+\Delta t)-p(a;t)}{\Delta t}=\sum_b [p(b;t)k_{b\to a}-p(a;t)k_{a\to b}]. \quad (104)$$

In the limit that $\Delta t\to 0$, we obtain

$$\dot{p}(a;t)=\sum_b [p(b;t)k_{b\to a}-p(a;t)k_{a\to b}]. \quad (105)$$

This is the master equation for single-particle trajectories (Gillespie, 1977; van Kampen, 1981; Zwanzig, 2001).

It is also straightforward to derive the evolution equation for a collection of M independent random walkers, which is an example of the chemical master equation. For simplicity, consider only two states A and B with $p_A(t)$ and $p_B(t)$ the probabilities of occupying states A and B at time t, respectively. The joint probability of having $M-n$ random walkers in state A and n random walkers in state B at time t is

$$P^{(M)}(M-n,n;t)\equiv \binom{M}{n} p_A(t)^{M-n}p_B(t)^n. \quad (106)$$

Using Eq. (101), we reexpress the populations $p_A(t)$ and $p_B(t)$ in terms of their populations $p_A(t-1)$ and $p_B(t-1)$ to derive an evolution equation. In this case, $P^{(M)}(M-n,n;t)$ is rewritten as

$$P^{(M)}(M-n,n;t)=\binom{M}{n}[p_A(t-1)(1-p_{A\to B}) + p_B(t-1)p_{B\to A}]^{M-n} \times [p_B(t-1)(1-p_{B\to A}) + p_A(t-1)p_{A\to B}]^n. \quad (107)$$

Taking the continuum limit once more by replacing $t-1$ by $t-\Delta t$, $p_{i\to j}$ by $\Delta t k_{i\to j}$, and letting $\Delta t\to 0$, we find

$$\dot{P}^{(M)}(M-n,n;t) = [-(M-n)k_{A\to B}-nk_{B\to A}]P^{(M)}(M-n,n;t) + (M-n+1)k_{A\to B}P^{(M)}(M-n+1,n-1;t) + (n+1)k_{B\to A}P^{(M)}(M-n-1,n+1;t), \quad (108)$$

which is the master equation for M random walkers (van Kampen, 1981; Zwanzig, 2001; Stock, Ghosh, and Dill, 2008). This expression is easily generalized to more than two states. When the walkers are not independent, and thus when the joint probabilities are not expressible in the simple form of Eq. (106), it is more difficult to compute the dynamical partition function, the sum over all allowed microtrajectories, analytically. However, the master equation can always be obtained from the discrete time evolution equation by keeping leading-order terms in the transition probability, just as we did in obtaining Eq. (108). Such modeling was applied to complexation (Ghosh, 2011) and can be applied to problems of self-catalysis, positive-feedback reactions including those having nonunit stoichiometric coefficients, and others.

The Fokker-Planck equation follows from the master equation in the limit of a large particle numbers (van Kampen, 1981; Zwanzig, 2001), or equivalently in the limit of small fluctuations. That is, discrete differences in particle numbers appearing on the right-hand side of the master equations are turned into derivatives with respect to particle numbers. This right-hand side of the Fokker-Planck equation is equivalent to the divergence of the flux of a diffusion equation. The flux landscape formulation of the Fokker-Planck equation (Wang, Xu, and Wang, 2008, 2009; Wang and Zaman, 2009; Wang, Li, and Wang, 2010; Wang *et al.*, 2010) has proven useful for explaining the robustness of biological clocks and oscillators in the presence of small noise (Wang, Xu, and Wang, 2008, 2009). See Haken (1986) for a generalization of the MaxCal treatment above to the case of a continuous state space and for a derivation of the Fokker-Planck equation from this continuous formalism; see Otten and Stock (2010) for a generalization of MaxCal incorporating time-dependent constraints.

X. NONEQUILIBRIUM STEADY STATES AND FLUCTUATION THEOREMS

There are many important relationships in nonequilibrium statistical mechanics, including the linear regression hypothesis, Onsager relations, the Green-Kubo relations, and fluctuation-dissipation theorems (Onsager, 1931a, 1931b; Casimir, 1945; Callen and Welton, 1951; Onsager and Machlup, 1953a, 1953b; Miller, 1956; Kubo, 1957; Kawasaki and Yamada, 1967; Jaynes and Rosenkrantz, 1989; Evans, Cohen, and Morriss, 1993; Evans and Searles, 1994, 2002; Gallavotti and Cohen, 1995a, 1995b; Luzzi, Vasconcellos, and Ramos, 2002; Fujisaki, Shiga, and Kidera, 2010). A recent focus of activity has been on fluctuation theorems (Evans, Cohen, and Morriss, 1993; Evans and Searles, 1994, 2002; Jarzynski, 1997; Crooks, 1999; Lebowitz and Spohn, 1999; Wang *et al.*, 2002; Dewar, 2003; van Zon and Cohen, 2003; Bodineau and Derrida, 2004, 2007; Seifert, 2005a, 2005b; Derrida, 2007; Harris and Schütz, 2007; Kurchan, 2007; Sevick *et al.*, 2008; Monthus, 2011). Dynamical quantities such as flux or current also were used to construct distributions and describe fluctuations away from equilibrium (Derrida and Lebowitz, 1998; Maes, 1999; Bodineau and Derrida, 2004; Depken and Stinchcombe, 2004; Bertini *et al.*, 2006; Bodineau and Derrida, 2007; Derrida, 2007; Hurtado and Garrido, 2009).

MaxCal gives insights into a flux fluctuation relationship (Monthus, 2011). Monthus (2011) showed that if the average flux $\langle J\rangle=\sum_C P_C J_C$ is a constraint, where J_C is the flux associated with a trajectory C, then the probability of that trajectory according to MaxCal is

$$P_C\sim \exp(\nu J_C), \quad (109)$$

where ν is the Lagrange multiplier that imposes the average flux constraint.

From large-deviation theory (LDT) (Ellis, 2006; Harris and Touchette, 2009; Touchette, 2009), we can assume the form of the flux distribution is

$$P(J) \sim \exp[-TI(j)], \tag{110}$$

where J is the integrated, or total, flux over an interval T and $j \equiv J/T$. For a brief summary of LDT see footnote.[10]

The form for $I(j)$, the "rate function" as it is called in large-deviation theory, depends on the microscopic details. For some problems, the detailed functional form of the large-deviation function $I(j)$ has been derived (Bodineau and Derrida, 2007).

To express the ratio of the probability of a forward trajectory P_C to a reverse trajectory P_{C_R}, we use Eq. (109),

$$\frac{P_C}{P_{C_R}} = \exp(2\nu J_C), \tag{113}$$

where we used the fact that $J_C = -J_{C_R}$ for the reverse path C_R.

The probability P_C for a specific microtrajectory is related to the probability $P(J)$ for observing flux J by a degeneracy factor $g(J)$. That is,

$$\begin{aligned} P(J) &= g(J)\exp(\nu J)/Q; \\ P(-J) &= g(-J)\exp(-\nu J)/Q, \end{aligned} \tag{114}$$

where $g(J)$ is the number of trajectories that have flux J, $g(-J)$ is the number of trajectories having flux $-J$, and Q is the dynamical partition function. For each forward trajectory, we have a reverse trajectory and their fluxes have opposite sign, so $g(J) = g(-J)$. This yields

$$P(J)/P(-J) = \exp(2\nu J). \tag{115}$$

Combining Eq. (110) with Eq. (115) gives

$$I(j) = I(-j) - 2\nu j. \tag{116}$$

Equation (116) is the fluctuation theorem for flux, as derived from MaxCal (Monthus, 2011). Similar results were derived from other approaches to path probabilities without explicitly using MaxCal (Maes, 1999; Bodineau and Derrida, 2007; Derrida, 2007). The advantage of the formulation above is that all the model specifics are contained within the rate function. Dewar has used similar arguments to derive a fluctuation theorem for entropy production (Dewar, 2003).

XI. MAXIMUM CALIBER IS USEFUL IN INTERPRETING EXPERIMENTS ON THE DYNAMICS OF FEW-PARTICLE SYSTEMS

In the following sections, we describe some particular applications of MaxCal in the analysis of experimental and simulated data.

A. Using MaxCal to describe diffusion in few-particle systems

In this section, we illustrate the application of MaxCal to few-particle diffusion. According to Fick's law, in macroscopic many-particle diffusion, a quantity of principle interest is the average flux $\langle J\rangle$, which is proportional to the concentration gradient. In one dimension we have

$$\langle J\rangle = -D\partial c/\partial x, \tag{117}$$

where D is the diffusion coefficient and c is the concentration. However, in few-particle diffusion, additional quantities are of interest but may not be accurately determinable from experiments. Such quantities include the variance of the flux $\langle J^2\rangle_c \equiv \langle J^2\rangle - \langle J\rangle^2$ that describes the more complete dynamical distribution function. MaxCal is useful for inferring quantities of this type from quantities that are measured. Experimental methods are now available that can test predictions about such fluctuational quantities. As a simple example, the flux fluctuations in the diffusion of small numbers of particles were explored in a microfluidics diffusion experiments performed by Rob Phillips and co-workers at Caltech (Ghosh *et al.*, 2006; Seitaridou *et al.*, 2007). Colloids were confined on one side of a gate [see Figs. 3(a) and 3(b)]. When the gate was opened, the particles diffused through the solvent in a tube and the location of the particles was tracked in time. This is essentially a diffusion experiment miniaturized to follow a small countable number of particles.

These experiments reveal more than just the average flux $\langle J\rangle$ of particles; from one cross-sectional column x compartment of the tube to the next, they monitor the full distribution of the jumps from one compartment to the next between neighboring time intervals, t and $t + \Delta t$. Imagine that there are N_1 beads in some thin cross-sectional slice, which we refer to as state 1, and N_2 in a neighboring cross-sectional slice on its right at time t, which we call state 2. In this problem, we consider pairwise constraints. There are no constraints on single-state occupancies. That is, we constrain transition probabilities between both states. Our dynamical partition function after one time step with a given initial condition,

$$\begin{pmatrix} a_1 & a_2 \end{pmatrix},$$

[10] If a coin is tossed N times and the outcome each time is $x_i = 0, 1$, where 0 is tails and 1 is heads, we can construct another random variable X_N, which is the average of these independent random variables,

$$X_N = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{111}$$

X_N's value lies between 0 and 1. In large-deviation theory we are interested in the probability distribution $P(X_N)$ of the outcomes. While much of probability theory focuses on small expansions near the peak of the distribution, $\sim P(X_N = 0.5)$ in this case (for example, near the point of 5000 heads out of 10 000 coin flips), LDT instead focuses on larger deviations further away from the peak (for example, around 7000 heads out of 10 000 coin flips) in the limit of large N. LDT shows on general grounds that $P(X_N)$ must have the form

$$P(X_N) \sim \exp[-Nf(X_N)], \tag{112}$$

where $f(x)$ is a function that has a minimum, the detailed form of which depends on the problem at hand. For example, $f(x) = \log 2 + x\log x + (1-x)\log(1-x)$ (Harris and Touchette, 2009) for the coin toss problem, with the minimum at $x = 1/2$.

takes the form of Eq. (90),

$$q = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} e^{-\lambda_{11}} & e^{-\lambda_{21}} \\ e^{-\lambda_{12}} & e^{-\lambda_{22}} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{118}$$

If the initial condition is

$$\begin{pmatrix} a_1 & a_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix},$$

then the microtrajectory starts from state 1. Alternatively, if

$$\begin{pmatrix} a_1 & a_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix},$$

then the microtrajectory starts from state 2. we denote by q_1 and q_2 the dynamical partition function for the microtrajectory starting from states 1 and 2. Since there is no drift in the fluid carrying the particles, we assume the following: $e^{-\lambda_{11}} = e^{-\lambda_{22}}$ and $e^{-\lambda_{12}} = e^{-\lambda_{21}}$. Also, for shorthand, we denote $p \equiv e^{-\lambda_{12}}/(e^{-\lambda_{11}} + e^{-\lambda_{12}})$. Knowing p is equivalent to knowing the diffusion constant D, which would also be needed in advance for any macroscopic study of diffusion.

For independent particles, the dynamical partition function for N_1 such particles starting from state 1 is

$$Q_1 = q_1^{N_1}. \tag{119}$$

The partition function in state 2 is similarly defined. For a single time step, the total dynamical partition function for N_1 particles in state 1 and N_2 in state 2 is therefore

$$Q = Q_1 Q_2. \tag{120}$$

Next we define the flux $J \equiv n_1 - n_2$, where n_1 is the stochastic number of particles going from state 1 to 2 and vice versa for n_2. We are interested in knowing the probability distribution of J predicted from MaxCal based on these simple pairwise constraints. This distribution is computed from

$$P(J) = Q'/Q, \tag{121}$$

where Q' is the restricted sum over those microtrajectory weights with net flux J while Q is the full dynamical partition function. The details of the calculation are in Pressé *et al.* (2010). The flux distribution then follows (Ghosh *et al.*, 2006; Seitaridou *et al.*, 2007):

$$P(J) = \frac{1}{\sqrt{2\pi N p(1-p)}} \exp\left(-\frac{[J - p(N_1 - N_2)]^2}{2Np(1-p)}\right), \tag{122}$$

where $N = N_1 + N_2$. Many properties can be inferred from the parametrized flux distribution. Figure 3 shows various experimental tests of the full flux distribution predicted in this way, including variances. The model also predicts so-called "bad actors," the small fraction of flows that, such as Maxwell's demon, flow up concentration gradients because of the small-numbers fluctuations, rather than down concentration gradients. Like fluctuation theorems (Evans, Cohen, and Morriss, 1993; Evans and Searles, 1994, 2002; Jarzynski, 1997; Crooks, 1999; Lebowitz and Spohn, 1999; Wang *et al.*, 2002; Dewar, 2003; van Zon and Cohen, 2003; Seifert, 2005a, 2005b; Bodineau and Derrida, 2007; Derrida, 2007; Harris and Schütz, 2007; Kurchan, 2007; Sevick *et al.*, 2008; Monthus, 2011), MaxCal makes predictions about the probabilities of rare events. Figures 3(c)–3(e) show that the

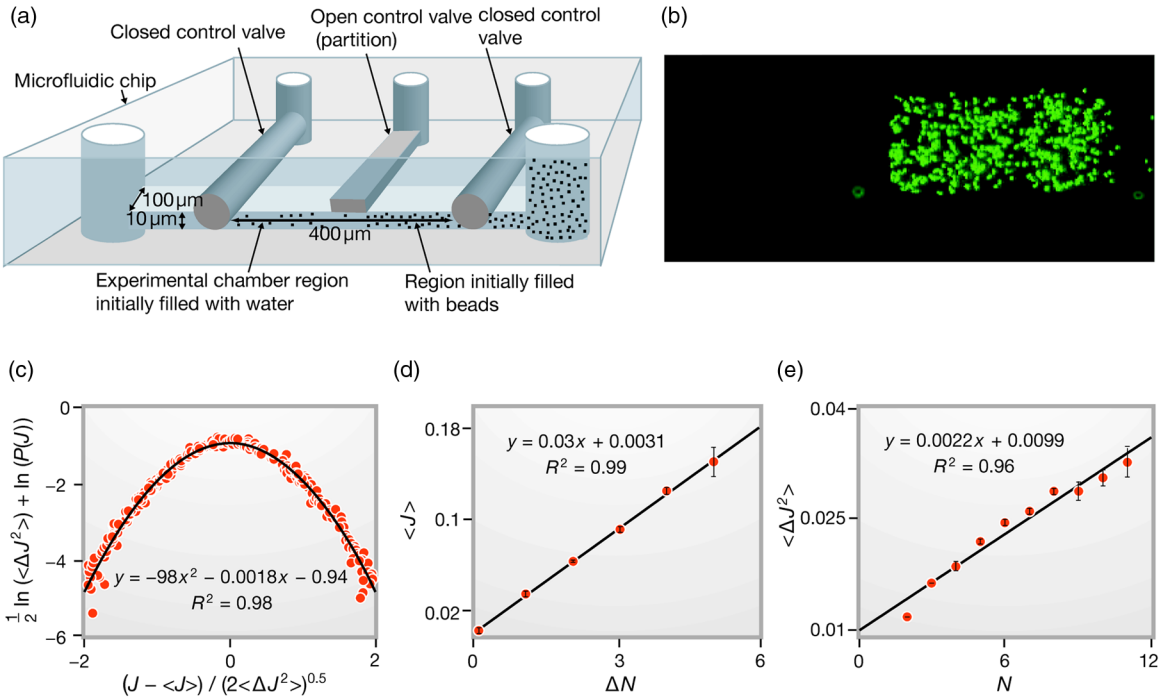

FIG. 3 (color online). Microfluidics experiment on few-particle to test MaxCal. Colloids corralled on one side of a gate begin to diffuse at time $t = 0$ when the gate is opened. (a) Schematic of the microfluidics chip. (b) Particle diffusion. From Seitaridou *et al.*, 2007. (c) The flux distribution function $\frac{1}{2}\ln[\langle(\Delta J)^2\rangle] + \log[P(J)]$ vs $(J - \langle J\rangle)/\sqrt{2\langle(\Delta J)^2\rangle}$. (d) The average flux $\langle J\rangle$ is shown as a function of $\Delta N = N_1 - N_2$. (e) The second cumulant $\langle\Delta J^2\rangle = \langle J^2\rangle - \langle J\rangle^2$ vs the total number of particles $N = N_1 + N_2$. For (c), (d), and (e), circles show experimental data; solid line shows MaxCal theory.

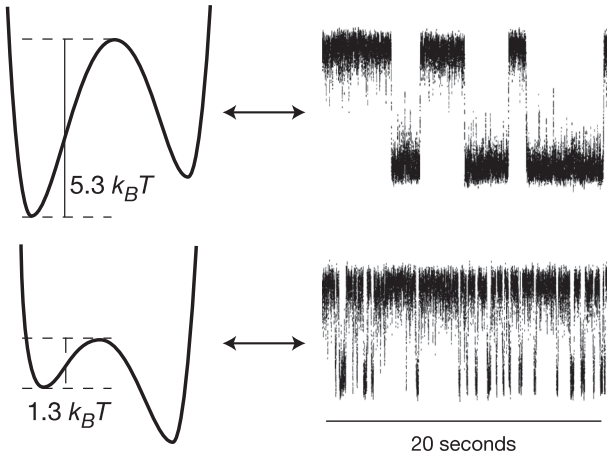

FIG. 4. A dual-laser-trap experiment in which a colloidal particle hops between two "sculpted" energy wells, back and forth, with observed time trajectories as shown. From Wu *et al.*, 2009.

agreement of the MaxCal inference with the experimental data is excellent.

B. Using MaxCal to describe single-particle two-state "reaction" kinetics

In the previous example, we considered a situation involving multiple independent particles. Here we describe the application of MaxCal to single-molecule two-state dynamics. Consider the dynamics of a two-state reaction between two states, 1 and 2, such as a chemical reaction, ion-channel opening or closing, or a biomolecule undergoing cycles of folding and unfolding; see Fig. 2. Rob Phillips and coworkers (Wu *et al.*, 2009) used dual laser traps to capture a single colloidal particle in one of two energy wells. They were able to control both the equilibrium and the kinetics by varying the well depths and the barrier height between the optical trap energies (Wu *et al.*, 2009). The bead hops between two wells with a stochastic duration time in each well. A time series of these hops defines the microtrajectories. A typical energy landscape and microtrajectories are shown in Fig. 4.

Our dynamical partition function is given by Eq. (90) and, as with the previous problem, we constrain our pairwise statistics $N_{1\to1}$, $N_{1\to2}$, $N_{2\to1}$, and $N_{2\to2}$. There are no singlet constraints, only the pairwise constraints are given.

The MaxCal strategy is to first parametrize the four Lagrange multipliers (λ_{ij} for $i,\ j=1,2$) from averages ($\langle N_{i\to j}\rangle$ for $i,\ j=1,2$) taken from the raw trajectories (Stock, Ghosh, and Dill, 2008; Otten and Stock, 2010). Again, only two Lagrange multipliers are independent from the normalization condition given in Eq. (76).

The Lagrange multipliers are parametrized through the following relation:

$$\langle N_{i\to j}\rangle = \frac{\partial \log Q}{\partial \lambda_{ij}} \tag{123}$$

by setting $\langle N_{i\to j}\rangle$ equal to the measured average for $i,\ j=1,2$. Once the Lagrange multipliers are determined, the dynamical partition function and microtrajectory probabilities are fully determined. With a parametrized dynamical partition function, the variance in transitions between sites can be inferred from relations such as

$$\langle (N_{1\to2})^2\rangle - \langle N_{1\to2}\rangle^2 = \frac{\partial^2 \log Q}{\partial \lambda_{1\to2}^2}. \tag{124}$$

Various higher cumulants were inferred in this way from the dynamical averages measured in experiment (Wu *et al.*, 2009). Figure 5 shows the good agreement between the MaxCal theory and the experiments.

C. MaxCal predicts far-from-equilibrium properties of multistate cycles, such as molecular motors

MaxCal has been used to analyze the spinning noise in biochemical cycles and motors that are far from equilibrium. Consider a cycle having s states (Fig. 6 shows a cycle with $s=3$ states). The cycle is driven to "spin" in a forward direction, for example, by adenosine triphosphate (ATP) hydrolysis. Examples of such cycles include the HSP90 chaperone protein complex, which assists in protein folding (Southworth and Agard, 2008; Mickler *et al.*, 2009); circadian clock circuits, driven by changing phosphorylation states

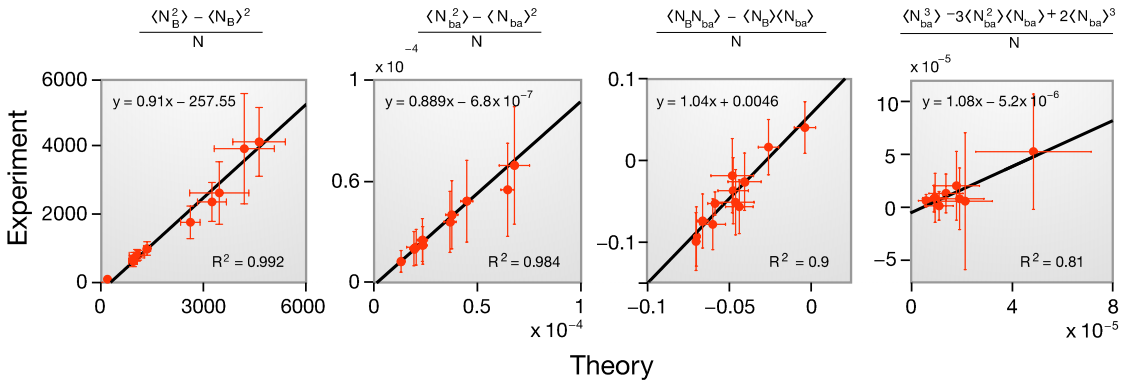

FIG. 5 (color online). The x axes give the predicted second cumulants, covariance, and third cumulants from the MaxCal approach, based on the known first moments. The y axes give the experimental values. The quantities are variance in N_1, $N_{1\to2}$ [which in the figure from the original source (Wu *et al.*, 2009) read N_B instead of N_1 and N_{ba} instead of $N_{1\to2}$], covariance of $N_1 N_{1\to2}$ and third cumulant of $N_{1\to2}$ from left to right. We define $N_1 = N_{1\to1} + N_{2\to1} + N_{0\to1}$, where $N_{0\to1} = \delta_{i_0,1}$. The dashed lines are the best linear fits; fitting parameters are inset. Each point represents one experimentally observed trajectory. Trajectories were 30 000 Δt units long, and errors were calculated for around 600 trajectories. From Wu *et al.*, 2009.

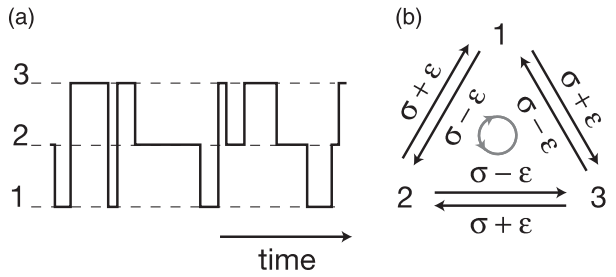

FIG. 6. (a) One example of a time trajectory for a three-state cycle. (b) A three-state cycle, showing the definitions of rates used in the text.

(Rust *et al.*, 2007); and Michaelis-Menten enzymes, which can be regarded as performing a three-state cycle, $E \leftrightarrow ES \leftrightarrow EP \leftrightarrow E$, driven to spin in a particular direction by high substrate concentrations.

Consider a simple cycle having two different types of pairwise constraints: pairwise constraints coinciding with forward motion along the cycle and backward motion along the cycle. In other words, we consider

$$e^{-\lambda_{mn}} \equiv \sigma + \epsilon; \qquad e^{-\lambda_{nm}} \equiv \sigma - \epsilon, \tag{125}$$

where m and n are neighboring states along the cycle. The Lagrange multiplier σ is related to an intrinsic rate that is the same in the forward and backward directions and ϵ represents some driving force, the degree to which detailed balance is broken to drive the system in one direction. In biology, ϵ may depend on the amount of ATP concentration driving a motor, for example. Given these two average rate quantities, MaxCal gives a way to infer the full distribution of dynamical trajectories including flux fluctuations (Pressé *et al.*, 2010). For this problem, MaxCal is complimentary to other methods (Qian and Elson, 2002; Seifert, 2005a, 2005b; Astumian, 2007; Andrieux and Gaspard, 2008), though the mathematics describing the dynamics are analogous to those of equilibrium statistical mechanics. Figure 7 shows the prediction that the fluctuational noise (for example, of occasional backward fluctuations) diminishes as the system is driven further away from equilibrium.

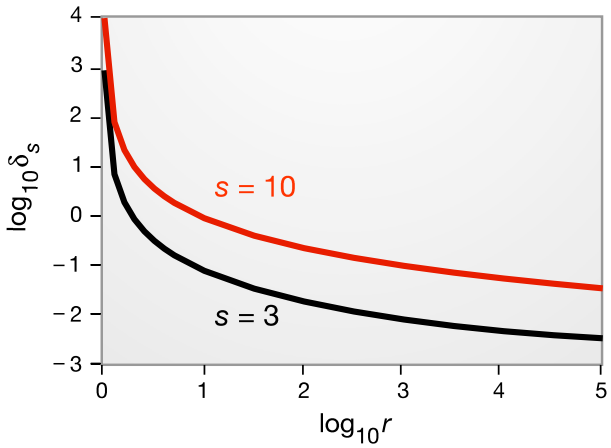

FIG. 7 (color online). A plot of $\delta_s = \langle J^2 \rangle_c / \langle J \rangle^2$ vs $r = [(\sigma+\epsilon)/(\sigma-\epsilon)]^2$, where s denotes the number of states in the cycle.

D. Path entropy maximization is useful for modeling neural spike trains

Schneidman *et al.* (2006) used a path entropy maximization procedure to infer a model for the firing patterns of arrays of neurons. They used the probability of a neuronal spike for each neuron as well as correlations in spiking behavior between different neurons within a small time interval δt, as their data constraints to build a model. They used a long-range Ising model. In the language we used earlier, we would speak of constraining single state occupancies (which microscopically coincide with the "fire" or "not fire" state of the neuron) as well as constraining single state occupancy correlations. With single state occupancy correlations they infer the probability of occurrence of a particular firing pattern for a set of 10 neurons (see Fig. 8), which would be impossible to predict by not constraining correlations. As an example of the model's predictive success, Schneidman *et al.* consider the probability of occurrence of the firing pattern 1011001010, where 1 stands for fired within $\delta t = 20$ ms and 0 otherwise for neurons 1 through 10. Schneidman *et al.* (2006) speak of the rate of a firing pattern as its probability of occurrence within the small time interval δt. They show that the rate of occurrence of the firing pattern 1011001010 is well captured by a long-range Ising model (predicting an average rate of occurrence of this pattern of about once a minute) while the independent Ising model errs by a factor of 10^6.

E. Application of MaxCal to a genetic toggle switch

For systems that have feedback, modeling few-particle dynamics is challenging because particles are coupled. Here we describe how MaxCal has been applied to a bistable system (Bagowski and Ferrell, Jr., 2001; Paliwal *et al.*, 2007; Raj and van Oudenaarden, 2008), specifically the synthetic genetic toggle switch of Gardner and co-workers (Gardner, Cantor, and Collins, 2000).

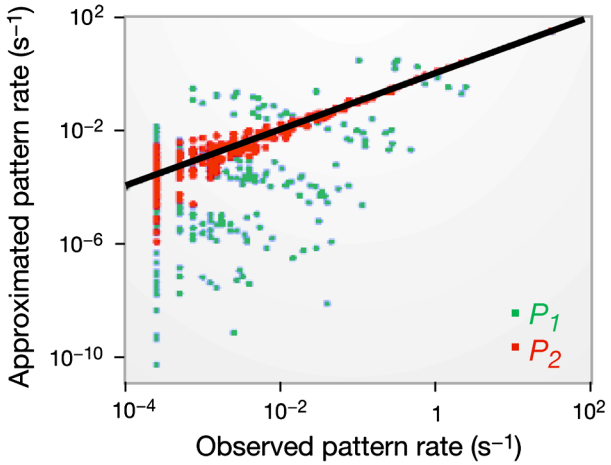

FIG. 8 (color online). Rate of occurrence of firing patterns with the approximated pattern rate vs the observed pattern rate. The solid black line indicates strict agreement. The (darker) colored dots coincide with the prediction made from the MaxEnt model where correlations were used as constraints, model P_2. The other (lighter) colored dots coincide with those predictions made from the independent Ising model, model P_1, where only mean firing patterns of each neuron are used as constraints.

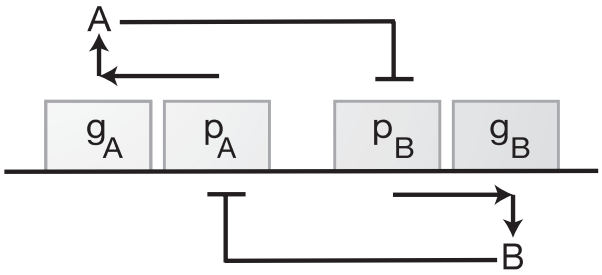

FIG. 9. The genetic toggle switch. The DNA plasmid is shown with promoters p_A and p_B of genes g_A and g_B that, when transcribed, produce proteins A and B, respectively. In this gene circuit, production of A inhibits production of B. And production of B inhibits A.

Figure 9 shows a cartoon of the gene circuit constructed by Gardner and co-workers (Gardner, Cantor, and Collins, 2000). The circuit encodes two proteins, A ("green") and B ("red"). It has two negative feedback loops. When green protein is made, red is suppressed, and when red protein is made, green is suppressed (see time trace in Fig. 10). The production of each protein is stochastic. An interesting aspect of bistable systems is two very different time scales. A rare fluctuation, which occurs on a fast time scale, can be strong enough to transition the circuit from one state to the other (green dominant or red dominant). The switch then remains stable in that state over much longer time scales. It is of interest to have a model that can explore how fluctuations can toggle the system between these stable states.

The process described in Fig. 9 has three dynamical components (Gardner, Cantor, and Collins, 2000; Lipshtat *et al.*, 2006). (1) Death processes: $A \rightarrow \phi$ indicates that A molecules are degraded with rate d_A; similarly for B. (2) Birth processes: $\alpha \rightarrow \alpha + A$ indicates that protein A is produced at rate k_α if the promoter α is active. Similarly $\beta \rightarrow \beta + B$ indicates that protein B is produced at rate k_β if the promoter β is active. (3) Binding events: $A + \beta \leftrightarrow \beta^*$ indicates that protein A converts B's active promoter β to inactive promoter β^* with rate f_A. The backward rate is r_A. Similarly, $B + \alpha \leftrightarrow \alpha^*$ indicates that protein B converts A's active promoter α to an inactive form α^* with rate f_B. The backward rate is r_B. Because the model treats only one gene, conservation requires that $[\alpha] + [\alpha^*] = 1$ and $[\beta] + [\beta^*] = 1$, where the brackets indicate the concentrations of the promoters. There is an additional constraint $[\alpha^*] + [\beta^*] \leq 1$ because of the system's design as an exclusive toggle switch. With the latter constraint, the only possible steady macroscopic states are (high A and low B) or (low A and high B),

$$\begin{aligned} &A \xrightarrow{d_A} \phi; \qquad \alpha \xrightarrow{k_\alpha} \alpha + A; \qquad B + \alpha \underset{r_B}{\overset{f_B}{\rightleftarrows}} \alpha^* \\ &B \xrightarrow{d_B} \phi; \qquad \beta \xrightarrow{k_\beta} \beta + B; \qquad A + \beta \underset{r_B}{\overset{f_B}{\rightleftarrows}} \beta^*. \end{aligned} \tag{126}$$

A model of the system dynamics was given by Gardner and co-workers (Gardner, Cantor, and Collins, 2000). They expressed the kinetics as a set of coupled differential equations using Hill-cooperativity terms to capture the bistable macroscopic state,

$$\begin{aligned} du/dt &= \alpha_1/(1 + v^\beta) - u; \\ dv/dt &= \alpha_2/(1 + u^\gamma) - v, \end{aligned} \tag{127}$$

where α_1 is the synthesis rate of the protein that has concentration $u(t)$ and α_2 is the synthesis rate of the protein having concentration $v(t)$. β and γ are adjustable Hill exponents. With this approach, however, (a) the mechanism must be known in advance [see Eq. (126)]; (b) the choices of nonlinear functional forms are arbitrary, and normally not obtainable by independent experiments; and (c) the fit parameters only guarantee agreement with average rates, not with the full rate distribution. No information about the fluctuations is obtained from such a model.

The MaxCal procedure is different. The dynamical partition function we construct here is a slight generalization of that given in Eq. (87). We have flux of production and degradation for each species. We impose these constraints by two binary indicator variables for each species; one for production and one for degradation. We let ℓ_α (or ℓ_β) be 0 when A (or B) is not produced within a small interval of time δt, or 1 otherwise. We choose the time interval δt to be sufficiently small that no more than a single A or B is

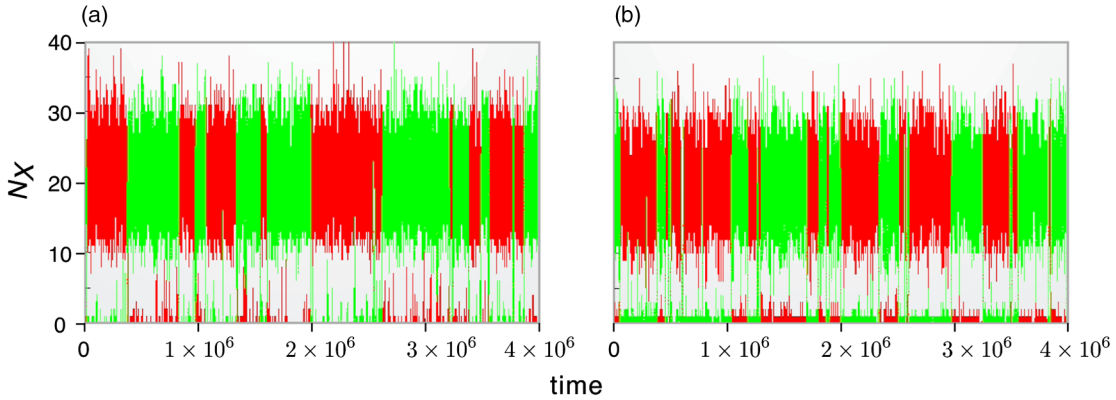

FIG. 10 (color online). Time trajectories of the toggle switch. (a) Gillespie simulation (Gillespie, 1977) of the model using $d = 0.005$, $k = 0.1$, $f = 100$, and $r = 2$. This acts as our "experimental data," with fully known underlying behavior. (b) MaxCal time trace of the model that we then extracted from the averages of the "experiment." This figure only validates that the parameters extracted by MaxCal do give trajectories resembling those from which they are extracted. The main quantitative tests, described in text and in the following figures, are the inference of entire statistical distributions.

produced within it. We assign variables $\ell_{i,A}$ and $\ell_{i,B}$ to each of the i individual protein molecules of type A and B, respectively. The variable $\ell_{i,A}$ or $\ell_{i,B}$ equals 1 when the ith particle is not degraded within interval δt, or 0 otherwise.

Thus, the dynamical partition function is

$$Q = \sum_{\substack{\ell_\alpha,\ell_\beta \\ \{\ell_A\},\{\ell_B\}}} \mathcal{Z}(\ell_\alpha, \ell_\beta, \{\ell_A\}, \{\ell_B\}) \exp\Big(h_\alpha \ell_\alpha + h_\beta \ell_\beta + h_A \sum_{i=1}^{N_A} \ell_{i,A} + h_B \sum_{i=1}^{N_B} \ell_{i,B}\Big). \tag{128}$$

The Lagrange multipliers (h_α, h_A) enforce the average observed production and degradation rates of protein A; similarly for B. The first two terms do not have summation because there is only one promoter for gene A and B. However, in our earlier discussion all particles were independent hence $\mathcal{Z}$ was assumed to be unity. In this example where one species regulates the other, we need to further impose constraints between different species. This is done by the function $\mathcal{Z}$ expressing the observed correlations between A and B,

$$\mathcal{Z} = \exp\Big(K_{A\beta} \sum_i \ell_\beta \ell_{i,A} + K_{B\alpha} \sum_i \ell_\alpha \ell_{i,B}\Big), \tag{129}$$

where $K_{A\beta}$ and $K_{B\alpha}$ are the correlation coupling parameters capturing the observation that high $[B]$ is associated with low $[A]$ and vice versa (Pressé, Ghosh, and Dill, 2011). Written in the form of a partition function, $Q = \sum \exp(-\mathcal{H})$ with a Hamiltonian-like quantity $\mathcal{H}$, we have

$$\mathcal{H} = h_\alpha \ell_\alpha + h_\beta \ell_\beta + h_A \sum_{i=1}^{N_A} \ell_{i,A} + h_B \sum_{i=1}^{N_B} \ell_{i,B} + K_{A\beta} \sum_i \ell_\beta \ell_{i,A} + K_{B\alpha} \sum_i \ell_\alpha \ell_{i,B}. \tag{130}$$

Our dynamical partition function resembles a long-range Ising model.

In short, in the MaxCal procedure the data trace is used directly to extract the Lagrange multipliers for production, degradation, and the correlations of the two species. Combined with the partition function, the data trace now gives the full dynamics [averages, fluctuations, higher cumulants, etc. (Pressé, Ghosh, and Dill, 2011)]. The quantities obtained in this way are not necessarily identical to those of a mass-action model having particular parameters, but they are more useful insofar as they do specify the full kinetic behavior of the system, they do so uniquely, and they do not require arbitrary invention of reaction topologies or functional forms in advance. This is evident by construction of an effective Hamiltonian, Eq. (130), that is independent of network topology and approximates the effects of the topology via the data used to constrain the model. Thus, MaxCal provides a powerful way to map a wide range of network topologies and reaction rates into an effective Hamiltonian-like function which is as reliable as the data used to constrain it. Combinations of different values of coupling constants explore the topology space such as negative or positive values of K's, corresponding to negative or positive feedback. The fluctuation spectrum which determines the statistics of switching from one state to another is inferred, not imposed by hand.

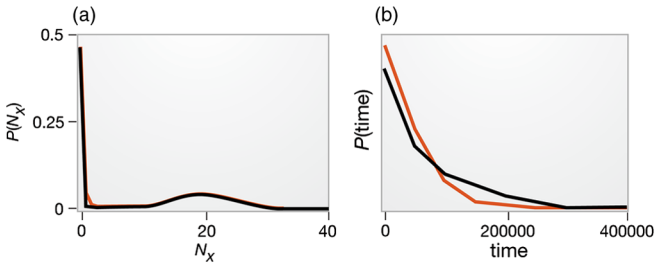

FIG. 11 (color online). Prediction of distributions by MaxCal and comparison with synthetic data. (a) Distribution of particle numbers. (b) Distribution of dwell times. Black (darker) curves: Gillespie "real data" simulations. Colored (lighter) curves: model predicted typical trajectory using the parameters extracted by MaxCal using Eq. (130). From Pressé, Ghosh, and Dill, 2011.

In Fig. 11 we show how MaxCal captures such nontrivial dynamical distributions quite accurately (Pressé, Ghosh, and Dill, 2011). Such distributions are not obtainable from mass-action models. MaxCal has other advantages: (1) it needs fewer parameters (3 for MaxCal vs 4 for the chemical master equation for the symmetric double negative feedback case, otherwise 6 for MaxCal and 8 for the asymmetric case), (2) the model is unique (only one model arises from given data), (3) much shorter time traces are required for MaxCal. Unlike other approaches, many switching events need not be observed to reasonably estimate the Lagrange multiplier values. And (4) the MaxCal model is extracted more directly from the data and less from an assumed model; degradation and synthesis events of A and B as well as their correlations are taken from direct observations over time. Given these observables, MaxCal builds a model without assuming functional forms or nonlinearities. It simply maps the direct observables into an effective dynamical Hamiltonian; see Eq. (130).

XII. CONCLUSIONS

We reviewed the principles of maximum entropy and maximum caliber. The quantity $H = -\sum p_j \log p_j$ has been used in variational principles across a variety of contexts. Boltzmann and Gibbs applied it to predicting material equilibria. Shannon used it to compute channel capacities for information transmission. Jaynes used it to frame physical problems as matters of inference.

We also reviewed the work of Shore and Johnson and others who showed more broadly in the 1980s that entropy maximization is a unique and sound recipe for ensuring basic consistency axioms when drawing inferences about distribution functions from observations. Entropy maximization is neither restricted to thermal equilibria of materials nor to matters of information transmission. Rather, it also provides a sound foundation for applications, illustrated here, ranging from image reconstruction to the inference of mathematical models from experimental data. We described the basis for the broad types of constraints that are allowable within such variational principles. First-moment constraints are common in physics, largely because physical systems commonly have

a large number of particles. But entropy maximization principles are not limited to those types of constraints; other constraints are relevant in image reconstruction, for example. We also discussed how entropy maximization sometimes gives a useful inversion of logic in modeling: rather than assuming a model and fitting its parameters, MaxEnt methods are sometimes used more directly to invert the data, with fewer model assumptions required up front.

We described maximum caliber in some detail, a MaxEnt-like variational principle that is useful for predicting physical dynamics, including for systems far from equilibrium. In maximum caliber, the populations p_j of dynamical pathways are predicted through a procedure of maximizing a route entropy subject to dynamical constraints, such as average rates or fluxes. We also showed that taking maximum caliber as a foundational principle gives insights into the roots of Markovian and master-equation approaches. Finally, we described several applications of MaxCal to single-molecule and few-particle dynamics experiments in biology and nanotechnology.

ACKNOWLEDGMENTS

We thank Ralph Chamberlin, Chris Fennell, Hao Ge, Justin MacCallum, Hong Qian, Ron Siegel, Gerhard Stock, and Attila Szabo for their thoughtful insights and comments. We appreciate the support of NIH Grant No. R01GM090205-03. S. P. acknowledges start-up support provided by IUPUI. K. G. acknowledges receiving research support from the Research Corporation for Science Advancement.

REFERENCES

Abe, S., 2000, Phys. Lett. A **271**, 74.
Abe, S., 2001, Phys. Rev. E **63**, 061105.
Aczél, J., 1966, *Lectures on Functional Equations and Their Applications* (Academic, New York).
Amari, S.-I., and H. Nagaoka, 2000, *Methods of Information Geometry* (Oxford University, New York).
Andrieux, D., and P. Gaspard, 2008, J. Chem. Phys. **128**, 154506.
Arkin, A., and J. Ross, 1995, J. Phys. Chem. **99**, 970.
Astumian, R. D., 2007, Proc. Natl. Acad. Sci. U.S.A. **104**, 19715.
Attard, P., 2009, Annu. Rep. Prog. Chem., Sect. C: Phys. Chem. **105**, 63.
Bagowski, C. P., J. E. Ferrell, Jr., 2001, Curr. Biol. **11**, 1176.
Balescu, R., 1975, *Equilibrium and Non-Equilibrium Statistical Mechanics* (Wiley, New York).
Balian, R., 2007, *From Microphysics to Macrophysics: Methods and Applications of Statistical Physics* (Springer-Verlag, Berlin), Vol. I.
Ben-Naim, A., 1985, *A Farewell to Entropy: Statistical Thermodynamics Based on Information* (World Scientific, Singapore).
Bertini, L., A. De Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim, 2006, J. Stat. Phys. **123**, 237.
Bodineau, T., and B. Derrida, 2004, Phys. Rev. Lett. **92**, 180601.
Bodineau, T., and B. Derrida, 2007, C.R. Physique **8**, 540.
Boltzmann, L., 1896, *Vorlesungen über Gastheorie, Teile I und II* (Verlag von Johann Ambrosius Barth, Leipzig).
Bratsun, D., D. Volfson, L. S. Tsimring, and J. Hasty, 2005, Proc. Natl. Acad. Sci. U.S.A. **102**, 14593.
Brush, S., 1975, *The Kind of Motion We Call Heat: A History of the Kinetic Theory of Gases in the 19th Century, Book 1* (North-Holland, Amsterdam)
Brush, S., 1976, *The Kind of Motion We Call Heat: A History of the Kinetic Theory of Gases in the 19th Century, Book 2* (North-Holland, Amsterdam)
Brush, S., 1983, *Statistical Physics and the Atomic Theory of Matter: From Boyle and Newton to Landau and Onsager* (Princeton University, Princeton, NJ).
Bryan, R. K., 1990, Eur. Biophys. J. **18**, 165.
Bustamante, C., W. Cheng, and Y. X. Mejia, 2011, Cell **144**, 480.
Çağatay, T., M. Turcotte, M. B. Elowitz, J. Garcia-Ojalvo, and G. M. Süel, 2009, Cell **139**, 512.
Callen, H. B., and T. A. Welton, 1951, Phys. Rev. **83**, 34.
Cao, J., and R. J. Silbey, 2008, J. Phys. Chem. B **112**, 12867.
Casimir, H. B. G., 1945, Rev. Mod. Phys. **17**, 343.
Caticha, A., and R. Preuss, 2004, Phys. Rev. E **70**, 046127.
Cecconi, C., E. A. Shank, C. Bustamante, and S. Marqusee, 2005, Science **309**, 2057.
Chamberlin, R. V., 1999, Phys. Rev. Lett. **82**, 2520.
Chamberlin, R. V., 2000, Nature (London) **408**, 337.
Chamberlin, R. V., 2002, Science **298**, 1172.
Chamberlin, R. V., J. V. Vermaas, and G. H. Wolf, 2009, Eur. Phys. J. B **71**, 1.
Chamberlin, R. V., and G. H. Wolf, 2009, Eur. Phys. J. B **67**, 495.
Chandler, D., 1987, *Introduction to Modern Statistical Mechanics* (Oxford University, New York).
Clausius, R., 1850a, Ann. Phys. (Berlin) **155**, 368.
Clausius, R., 1850b, Ann. Phys. (Berlin) **155**, 500.
Cox, R. P., 1946, Am. J. Phys. **14**, 1.
Cox, R. T., 1961, *The Algebra of Probable Inference* (Johns Hopkins Press, Baltimore).
Crooks, G. E., 1999, Phys. Rev. E **60**, 2721.
Csiszár, I., 1991, Ann. Stat. **19**, 2032.
de Groot, S. R., and S. R. Mazur, 1962, *Non-Equilibrium Thermodynamics* (North-Holland, Amsterdam).
Denbigh, K. G., and J. S. Denbigh, 1985, *Entropy in Relation to Incomplete Knowledge* (Cambridge University Press, Cambridge, England).
Depken, M., and R. Stinchcombe, 2004, Phys. Rev. Lett. **93**, 040602.
Derrida, B., 2007, J. Stat. Mech. P07023.
Derrida, B., and J. L. Lebowitz, 1998, Phys. Rev. Lett. **80**, 209.
Dewar, R., 2003, J. Phys. A **36**, 631.
Dewar, R., 2005, J. Phys. A **38**, L371.
Dewar, R. C., 2009, Entropy **11**, 931.
Dill, K. A., and S. Bromberg, 2011, *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology* (Garland Science, New York).
Eldar, A., and M. B. Elowitz, 2010, Nature (London) **467**, 167.
Ellis, R. S., 2006, *Entropy, Large Deviations, and Statistical Mechanics* (Springer-Verlag, Berlin).
Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain, 2002, Science **297**, 1183.
El Samad, H., M. Khammash, L. Petzold, and D. Gillespie, 2005, Int. J. Robust Nonlinear Control **15**, 691.
Engl, W., K. Kunisch, and A. Neubauer, 1989, Inverse Probl. **5**, 523.
Evans, D. J., E. G. D. Cohen, and G. P. Morriss, 1993, Phys. Rev. Lett. **71**, 2401.
Evans, D. J., and D. J. Searles, 1994, Phys. Rev. E **50**, 1645.
Evans, D. J., and D. J. Searles, 2002, Adv. Phys. **51**, 1529.
Evans, R. M. L., 2004a, Phys. Rev. Lett. **92**, 150601.
Evans, R. M. L., 2004b, Physica (Amsterdam) **340A**, 364.

Evans, R. M. L., 2005, J. Phys. A **38**, 293.
Feinstein, A., 1958, *Foundations of Information Theory* (McGraw-Hill, New York).
Filyukov, A. A., 1968, J. Eng. Phys. Thermophys. **14**, 814.
Filyukov, A. A., and V. Y. Karpov, 1967a, J. Eng. Phys. **13**, 624.
Filyukov, A. A., and V. Y. Karpov, 1967b, J. Eng. Phys. **13**, 798.
Flomenbom, O., J. Klafter, and A. Szabo, 2005, Biophys. J. **88**, 3780.
Flomenbom, O., and R. J. Silbey, 2006, Proc. Natl. Acad. Sci. U.S.A. **103**, 10907.
Fowler, R. H., 1938, *Statistical Mechanics* (Cambridge University Press, Cambridge, England).
Fujisaki, H., M. Shiga, and A. Kidera, 2010, J. Chem. Phys. **132**, 134101.
Gallavotti, G., and E. G. D. Cohen, 1995a, J. Stat. Phys. **80**, 931.
Gallavotti, G., and E. G. D. Cohen, 1995b, Phys. Rev. Lett. **74**, 2694.
Gardner, T. S., C. R. Cantor, and J. J. Collins, 2000, Nature (London) **403**, 339.
Gaspard, P., 2004, J. Stat. Phys. **117**, 599.
Ge, H., S. Pressé, K. Ghosh, and K. A. Dill, 2012, J. Chem. Phys. **136**, 064108.
Ghosh, K., 2011, J. Chem. Phys. **134**, 195101.
Ghosh, K., K. A. Dill, E. Seitaridou, M. Inamdar, and R. Phillips, 2006, Am. J. Phys. **74**, 123.
Gibbs, J. W., 1902, *Elementary Principles in Statistical Mechanics* (Yale University, New Haven, CT).
Gillespie, D. T., 1977, J. Phys. Chem. **81**, 2340.
Gull, S. F., and G. J. Daniell, 1978, Nature (London) **272**, 686.
Gull, S. F., and J. Skilling, 1989, Eds., *Maximum Entropy and Bayesian Methods*, "Developments in Maximum Entropy Data Analysis" (Kluwer Academic, Dordrecht), pp. 53–74.
Haken, H., 1986, Z. Phys. B **63**, 505.
Hamill, O. P., A. Marty, E. Neher, B. Sakmann, and F. J. Sigworth, 1981, Pfluegers Arch. **391**, 85.
Harris, R. J., and G. M. Schütz, 2007, J. Stat. Mech. P07020.
Harris, R. J., and H. Touchette, 2009, J. Phys. A **42**, 342001.
Hill, T. L., 1956, *Statistical Mechanics: Principles and Selected Applications* (McGraw-Hill, New York).
Hill, T. L., 1962, J. Chem. Phys. **36**, 3182.
Hill, T. L., 2001a, Nano Lett. **1**, 111.
Hill, T. L., 2001b, Nano Lett. **1**, 273.
Hille, B., 1994, Trends Neurosci. **17**, 531.
Huang, S., Y. P. Guo, G. May, and T. Enver, 2007, Dev. Biol. **305**, 695.
Hurtado, P. I., and P. L. Garrido, 2009, J. Stat. Mech. P02032.
Jarzynski, C., 1997, Phys. Rev. Lett. **78**, 2690.
Jaynes, E. T., 1957a, Phys. Rev. **106**, 620.
Jaynes, E. T., 1957b, Phys. Rev. **108**, 171.
Jaynes, E. T., 1968, IEEE Trans. Sys. Sci. Cyber. **4**, 227.
Jaynes, E. T., 2003, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, England).
Jaynes, E. T., and K. Ford, 1963, Eds., *Information Theory and Statistical Mechanics—Brandeis University Summer Institute Lectures in Theoretical Physics* (Benjamin, New York).
Jaynes, E. T., and H. Haken, 1985, Eds., *Complex Systems-Operational Approaches* (Springer-Verlag, Berlin).
Jaynes, E. T., R. D. Levine, and M. Tribus, 1979, Eds., *The Maximum Entropy Formalism* (MIT, Cambridge, MA).
Jaynes, E. T., and S. Rabinovitch, 1980, Eds., *Annual Review of Physical Chemistry* (Annual Reviews, Palo Alto).
Jaynes, E. T., and R. D. Rosenkrantz, 1989, Eds., *Papers on Probability, Statistics and Statistical Physics* (D. Reidel, Dordrecht).
Jeffreys, H., 1946, Proc. R. Soc. A **186**, 453.
Jeffreys, H., 1948, *Theory of Probability* (Oxford University, New York).
Kaern, M., T. C. Elston, W. J. Blake, and J. J. Collins, 2005, Nat. Rev. Genet. **6**, 451.
Kawasaki, K., and T. Yamada, 1967, Prog. Theor. Phys. **38**, 1031.
Kou, S. C., B. J. Cherayil, W. Min, B. P. English, and S. X. Xie, 2005, J. Phys. Chem. B **109**, 19 068.
Kubo, R., 1957, J. Phys. Soc. Jpn. **12**, 570.
Kullback, S., and R. A. Leibler, 1951, Ann. Math. Stat. **22**, 79.
Kurchan, J., 2007, J. Stat. Mech. P07005.
Landau, L., and E. M. Lifshitz, 1951, *Statistical Physics* (Mir, Moscow).
Landsberg, P. T., 1972, Nature (London) **238**, 229.
Landsberg, P. T., 1984, J. Stat. Phys. **35**, 159.
Lebowitz, J. L., and H. Spohn, 1999, J. Stat. Phys. **95**, 333.
Lecomte, V., C. Appert-Rolland, and F. van Wijland, 2007, J. Stat. Phys. **127**, 51.
Lee, J., and S. Pressé, 2012a, J. Chem. Phys. **137**, 074103.
Lee, J., and S. Pressé, 2012b, Phys. Rev. E **86**, 041126.
Lezon, T. R., J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Federoff, 2006, Proc. Natl. Acad. Sci. U.S.A. **103**, 19 033.
Lindley, D., 2001, *Boltzmann's Atom: The Great Debate That Launched a Revolution in Physics* (Free Press, New York).
Liphardt, J., B. Onoa, S. B. Smith, I. Tinoco, and C. Bustamante, 2001, Science **292**, 733.
Lipshtat, A., A. Loinger, N. Q. Balaban, and O. Biham, 2006, Phys. Rev. Lett. **96**, 188101.
Livesey, A. K., and J. C. Brochon, 1987, Biophys. J. **52**, 693.
Livesey, A. K., and J. Skilling, 1985, Acta Crystallogr. Sect. A **41**, 113.
Locasale, J. W., and A. Wolf-Yadlin, 2009, PLoS ONE **4**, e6522.
Lu, H. P., L. Xun, and X. S. Xie, 1998, Science **282**, 1877.
Luzzi, R., A. R. Vasconcellos, and J. G. Ramos, 2002, *Predictive Statistical Mechanics: A Nonequilibrium Ensemble Formalism* (Kluwer Academic, Boston)
Maes, C., 1999, J. Stat. Phys. **95**, 367.
McKinney, S. A., C. Joo, and T. Ha, 2006, Biophys. J. **91**, 1941.
McQuarrie, D. A., 2000, *Statistical Mechanics* (University Science, Sausalito).
Meinel, E. S., 1988, J. Opt. Soc. Am. A **5**, 25.
Methfessel, C., V. Witzemann, T. Takahashi, M. Mishina, S. Numa, and B. Sakmann, 1986, Pflügers Arch. **407**, 577.
Mickler, M., M. Hessling, C. Ratzke, J. Buchner, and T. Hugel, 2009, Nat. Struct. Mol. Biol. **16**, 281.
Milescu, L. S., G. Akk, and F. Sachs, 2005, Biophys. J. **88**, 2494.
Miller, D. G., 1956, Am. J. Phys. **24**, 433.
Moffitt, J. R., Y. R. Chemla, K. Aathavan, S. Grimes, P. J. Jardine, D. L. Anderson, and C. Bustamante, 2009, Nature (London) **457**, 446.
Monthus, C., 2011, J. Stat. Mech. P03008.
Munsky, B., B. Trinh, and M. Khammash, 2009, Mol. Syst. Biol. **5**, 318.
Onsager, L., 1931a, Phys. Rev. **37**, 405.
Onsager, L., 1931b, Phys. Rev. **38**, 2265.
Onsager, L., and S. Machlup, 1953a, Phys. Rev. **91**, 1505.
Onsager, L., and S. Machlup, 1953b, Phys. Rev. **91**, 1505.
Otten, M., and G. Stock, 2010, J. Chem. Phys. **133**, 034119.
Paliwal, S., P. A. Iglesias, K. Campbell, Z. Hilioti, A. Groisman, and A. Levchenko, 2007, Nature (London) **446**, 46.
Pathria, R. K., 1996, *Statistical Mechanics* (Butterworth-Heineman, Oxford, England).
Paulsson, J., 2004, Nature (London) **427**, 415.
Paulsson, J., 2005, Phys. Life Rev. **2**, 157.

Phillips, R., J. Kondev, and J. Theriot, 2009, *Physical Biology of the Cell* (Garland Science, New York).
Pressé, S., K. Ghosh, K. A. Dill, 2011, J. Phys. Chem. B **115**, 6202.
Pressé, S., K. Ghosh, R. Phillips, and K. A. Dill, 2010, Phys. Rev. E **82**, 031905.
Qian, H., and E. L. Elson, 2002, Biophys. Chem. **101–102**, 565.
Qin, F., A. Auerbach, and F. Sachs, 1997, Proc. R. Soc. B **264**, 375.
Qin, F., A. Auerbach, and F. Sachs, 2000, Biophys. J. **79**, 1915.
Raj, A., and A. van Oudenaarden, 2008, Cell **135**, 216.
Rao, C. V., and A. Arkin, 2003, J. Chem. Phys. **118**, 4999.
Rhoades, E., M. Cohen, B. Schuler, and G. Haran, 2004, J. Am. Chem. Soc. **126**, 14686.
Ross, J., 2003, Acc. Chem. Res. **36**, 839.
Rust, M. J., J. S. Markson, W. S. Lane, D. S. Fisher, and E. K. O'Shea, 2007, Science **318**, 809.
Samoilov, M., S. Plyasunov, and A. P. Arkin, 2005, Proc. Natl. Acad. Sci. U.S.A. **102**, 2310.
Sanchez, A., and J. Kondev, 2008, Proc. Natl. Acad. Sci. U.S.A. **105**, 5081.
Schneidman, E., M. J. Berry, R. Segev, and W. Bialek, 2006, Nature (London) **440**, 1007.
Schnitzer, M. J., and S. M. Block, 1995, Cold Spring Harbor Symposia on Quantitative Biology **60**, 793.
Schultz, D., E. Ben Jacob, J. N. Onuchic, and P. G. Wolynes, 2007, Proc. Natl. Acad. Sci. U.S.A. **104**, 17582.
Seifert, U., 2005a, Europhys. Lett. **70**, 36.
Seifert, U., 2005b, Phys. Rev. Lett. **95**, 040602.
Seitaridou, E., M. Inamdar, R. Phillips, K. Ghosh, and K. Dill, 2007, J. Phys. Chem. B **111**, 2288.
Sengupta, P., K. Garai, N. Balaji, N. Periasamy, and S. Maiti, 2003, Biophys. J. **84**, 1977.
Sevick, E. M., R. Prabhakar, S. R. Williams, and D. J. Searles, 2008, Annu. Rev. Phys. Chem. **59**, 603.
Shahrezaei, V., and P. S. Swain, 2008, Curr. Opin. Biotechnol. **19**, 369.
Shannon, C. E., 1948, Bell Syst. Tech. J. **27**, 379.
Shore, J. E., and R. W. Johnson, 1980, IEEE Trans. Inf. Theory **26**, 26.
Shore, J. E., and R. W. Johnson, 1981, IEEE Trans. Inf. Theory **27**, 472.
Siemiarczuk, A., B. D. Wagner, and R. W. Ware, 1990, J. Phys. Chem. **94**, 1661.
Skilling, J., 1984, Nature (London) **309**, 748.
Skilling, J., and R. K. Bryan, 1984, Mon. Not. R. Astron. Soc. **211**, 111.
Skilling, J., G. J. Erickson, and C. R. Smith, 1988, Eds., *Maximum-Entropy and Bayesian Methods in Science and Engineering* (Kluwer, Dordrecht)
Skilling, J., and S. F. Gull, 1991, IMS Lecture Notes: Monograph Series, Spatial Statistics and Imaging **20**, 341.
Smith, E., 2011, Rep. Prog. Phys. **74**, 046601.
Southworth, D. R., and D. A. Agard, 2008, Mol. Cell **32**, 631.
Steinbach, P. J., K. Chu, H. Frauenfelder, J. B. Johnson, D. C. Lamb, G. U. Nienhaus, T. B. Sauke, and R. D. Young, 1992, Biophys. J. **61**, 235.
Steinbach, P. J., R. Ionescu, and C. R. Matthews, 2002, Biophys. J. **82**, 2244.
Stock, G., K. Ghosh, and K. A. Dill, 2008, J. Chem. Phys. **128**, 194102.
Tikochinsky, Y., N. Z. Tishby, and R. D. Levine, 1984, Phys. Rev. Lett. **52**, 1357.
Tisza, L., 1963, Rev. Mod. Phys. **35**, 151.
Tisza, L., and P. M. Quay, 1963, Ann. Phys. (N.Y.) **25**, 48.
Tkacik, G., A. M. Walczak, and W. Bialek, 2009, Phys. Rev. E **80**, 031920.
Tolman, R. C., 1938, *The Principles of Statistical Mechanics* (Oxford University, New York).
Touchette, H., 2009, Phys. Rep. **478**, 1.
Tsallis, C., 1988, J. Stat. Phys. **52**, 479.
Tsallis, C., S. Abe, and Y. Okamoto, 2001, Eds., *Nonextensive Statistical Mechanics and Its Applications* (Springer-Verlag, Berlin)
Tsallis, C., M. Gell-Mann, and Y. Sato, 2005, Proc. Natl. Acad. Sci. U.S.A. **102**, 15377.
van Kampen, N. G., 1981, *Stochastic Processes in Chemistry and Physics* (North Holland, Amsterdam).
van Zon, R., and E. G. D. Cohen, 2003, Phys. Rev. Lett. **91**, 11601.
Vergassola, M., E. Villermaux, and B. I. Shraiman, 2007, Nature (London) **445**, 406.
Walczak, A. M., G. Tkacik, and W. Bialek, 2010, Phys. Rev. E **81**, 041905.
Wang, G. M., E. M. Sevick, E. Mittag, D. J. Searles, and J. Denis, 2002, Phys. Rev. Lett. **89**, 050601.
Wang, J., C. H. Li, and E. K. Wang, 2010, Proc. Natl. Acad. Sci. U.S.A. **107**, 8195.
Wang, J., L. Xu and E. Wang, 2009, Biophys. J. **97**, 3038.
Wang, J., L. Xu, and E. K. Wang, 2008, Proc. Natl. Acad. Sci. U.S.A. **105**, 12271.
Wang, J., L. Xu, E. K. Wang, and S. Huang, 2010, Biophys. J. **99**, 29.
Wang, J., and M. H. Zaman, 2009, Eds., *Potential Landscape Theory of Cellular Networks, Statistical Mechanics of Cellular Systems and Processes* (Cambridge University Press, Cambridge, England).
Warren, P. B., and P. R. ten Wolde, 2005, J. Phys. Chem. B **109**, 6812.
Watkins, L. P., H. Chang, and H. Yang, 2006, J. Phys. Chem. A **110**, 5191.
Witkoskie, J. B, and J. Cao, 2004, J. Chem. Phys. **121**, 6361.
Witkoskie, J. B, and J. Cao, 2008, J. Phys. Chem. B **112**, 5988.
Wright, P. G., 1970, Proc. R. Soc. A **317**, 477.
Wu, D., K. Ghosh, M. Inamdar, H. J. Lee, S. Fraser, K. A. Dill, and R. Phillips, 2009, Phys. Rev. Lett. **103**, 050603.
Yang, H., G. Luo, P. Karnchanaphanurach, T.-M. Louie, I. Rech, S. Cova, L. Xun, and S. X. Xie, 2003, Science **302**, 262.
Yang, H., and S. X. Xie, 2002, J. Chem. Phys. **117**, 10 965.
Yu, J., J. Moffitt, C. L. Hetherington, C. Bustamante, and G. Oster, 2010, J. Mol. Biol. **400**, 186.
Zubarev, D. N., 1961, Dokl. Akad. Nauk SSSR **140**, 92 [1962, Sov. Phys. Dokl. **6**, 776].
Zubarev, D. N., 1971, *Non-Equilibrium Statistical Thermodynamics* (Nauka, Moscow).
Zwanzig, R., 2001, *Nonequilibrium Statistical Mechanics* (Oxford University, New York)