# Conjugate variables in continuous maximum-entropy inference

Sergio Davis* and Gonzalo Gutiérrez†

*Grupo de Nanomateriales, Departamento de Física, Facultad de Ciencias, Universidad de Chile, Casilla 653, Santiago, Chile*

(Received 8 June 2012; revised manuscript received 20 July 2012; published 28 November 2012)

For a continuous maximum-entropy distribution (obtained from an arbitrary number of simultaneous constraints), we derive a general relation connecting the Lagrange multipliers and the expectation values of certain particularly constructed functions of the states of the system. From this relation, an estimator for a given Lagrange multiplier can be constructed from derivatives of the corresponding constraining function. These estimators sometimes lead to the determination of the Lagrange multipliers by way of solving a linear system, and, in general, they provide another tool to widen the applicability of Jaynes's formalism. This general relation, especially well suited for computer simulation techniques, also provides some insight into the interpretation of the hypervirial relations known in statistical mechanics and the recently derived microcanonical dynamical temperature. We illustrate the usefulness of these new relations with several applications in statistics.

DOI: 10.1103/PhysRevE.86.051136 PACS number(s): 05.20.−y, 05.70.−a, 02.50.Tt, 02.50.Cw

## I. INTRODUCTION

Statistical mechanics aims at explaining the physical properties of a macroscopic system from the basis of the dynamical properties of its microscopical constituents, for this relying strongly on the foundations of statistics.Several interpretations of the formalism produced by statistical mechanics have arisen. In particular, a Bayesian interpretation of statistical mechanics (and therefore of thermodynamics) was proposed by Jaynes [1] in 1957. In this interpretation, the probability assigned to a microstate $\vec{x}$ represents our degree of belief attached to the logical proposition "the physical system is in the microstate $\vec{x}$" and the information entropy $S(F_1,\ldots,F_m)$ quantifies our missing knowledge about the unknown microstate, given $m$ macroscopic properties $F_1,\ldots,F_m$. Maximization of the entropy subject to agreement with macroscopic observables leads naturally to the canonical, grand-canonical, and microcanonical probability distributions.

The virtue of the Bayesian viewpoint is the realization of the following: The maximum entropy (MaxEnt) formalism is a separate entity from thermodynamics; it is the most unbiased procedure for assigning degrees of belief to propositions, while being compatible with testable information. MaxEnt is, therefore, applicable in any branch of science where there is not enough information to decide among different competing hypotheses [2]. Moreover, the rationale of MaxEnt does not need to rely on its lack of bias (which might appear as a "subjective" quality), as the formalism has been recently recovered [3] only from axioms of logical consistency. It is not even necessary to invoke the concept of missing information to arrive at the same formalism [4].

Although several advances have been made in the fields of Bayesian inference since the seminal work of Jaynes [1] to the present day, there are still many aspects to explore, particularly in the case of a continuous maximum entropy distribution. For example, a recurrent problem is the evaluation of the Lagrange multipliers appearing in the Shannon-Jaynes entropy maximization procedure, which amounts to solving a system of nonlinear equations and, in most cases, has to be done numerically.

In this paper, we introduce a new identity (to the best of our knowledge) relating the expectation value of an arbitrary trial function to the values of Lagrange multipliers appearing in a MaxEnt continuous inference problem. This identity allows us to derive, for instance, the so-called hypervirial relations [5] of statistical mechanics as well as the recently proposed dynamical temperature [6,7], for the microcanonical ensemble, as particular cases, revealing the physics behind it. The dynamical temperature, a microscopic expression for the temperature of a continuous Hamiltonian, is now widely used in computer simulations of liquids [8] and proteins [9].

The paper is organized as follows. In Sec. II we review the MaxEnt formalism with continuous degrees of freedom. In Sec. III we derive a general relation which can be used to construct estimators of the Lagrange multipliers involved in continuous MaxEnt. In Sec. IV we recover several known relations from statistical mechanics, as well as others, as far as we know, not previously reported. Section V applies the relations derived in this work to MaxEnt problems within and beyond statistical mechanics.

## II. THE CONTINUOUS INFERENCE PROBLEM

Consider a set of $N$ variables or degrees of freedom $(x_1,\ldots,x_N)$ (denoted collectively by $\vec{x}$) subjected to uncertainty according to an initially assigned probability distribution $\pi(\vec{x})$ (the "complete ignorance" prior, in Bayesian terms). Suppose we want to revise this probability in light of new testable information, given in the form of $m$ known expectation values $\langle f_i(\vec{x})\rangle = F_i$ (where $i = 1,\ldots,m$) finally producing a new probability distribution $P(\vec{x})$. The updating procedure involves the maximization of the Shannon-Jaynes entropy (which is the negative of the Kullback-Leibler divergence [10]),

$$S[P(\vec{x}),\pi(\vec{x})] = -\int d\vec{x}\, P(\vec{x}) \ln \frac{P(\vec{x})}{\pi(\vec{x})}, \tag{1}$$

under the $m$ constraints, which can be written compactly as a single vector constraint, $\langle \vec{f}(\vec{x})\rangle = \vec{F}$.

*sdavis@gnm.cl
†gonzalo@fisica.ciencias.uchile.cl

 

The unique solution to this problem is

$$P(\vec{x}) = \frac{1}{\mathcal{Z}(\vec{\lambda})} e^{-\vec{\lambda}\cdot\vec{f}(\vec{x})}\pi(\vec{x}), \tag{2}$$

where the partition function $\mathcal{Z}$ is given by

$$\mathcal{Z}(\vec{\lambda}) = \int d\vec{x}\pi(\vec{x})e^{-\vec{\lambda}\cdot\vec{f}(\vec{x})} \tag{3}$$

and the vector of Lagrange multipliers $\vec{\lambda} = (\lambda_1, \ldots, \lambda_m)$ is implicitly given as the solution of

$$\vec{F} = -\frac{\partial}{\partial\vec{\lambda}} \ln \mathcal{Z}(\vec{\lambda}). \tag{4}$$

Alternatively, $\vec{\lambda}$ can be also obtained [2] from

$$\vec{\lambda} = \frac{\partial S(\vec{F})}{\partial \vec{F}}, \tag{5}$$

which can be recognized in the case of thermodynamics as the "definition" of temperature, $\partial S/\partial E = \beta$.

Note that the symmetry between Eqs. (4) and (5) is due to the fact that $\ln \mathcal{Z}$ and $S$ are connected via a Legendre transformation,

$$S = \ln \mathcal{Z} + \vec{\lambda}\cdot\vec{F}. \tag{6}$$

In this sense, the components of $\vec{\lambda}$ are conjugate to the corresponding components of $\vec{F}$. In thermodynamics, such pairs are common: For instance, internal energy $U$ and temperature $\beta$, volume $V$ and pressure $P$, number of particles $N$ and chemical potential $\mu$.

Equation (5), however, is usually impractical in analytical terms, as it requires knowledge of the entropy $S$ as a function of the known expectation values. It would be extremely practical to construct estimators $\hat{\lambda}_k(\vec{x_1}, \ldots, \vec{x_n})$ for the Lagrange multipliers $\lambda_k$ in order to avoid solving Eqs. (4) or (5) explicitly. Note, however, that this assumes we have access to a sample of $n$ points $\vec{x}$ (microstates, in the context of statistical mechanics) drawn from the original distribution. Fortunately, such a situation occurs frequently in the context of standard computer simulation techniques such as Monte Carlo and molecular dynamics.

Better yet, the existence of an estimator $\hat{\lambda}_k$ with the particular form $\langle\hat{\lambda}_k(\vec{x})\rangle = \lambda_k$ puts the Lagrange multiplier $\lambda_k$ as an *observable* (that is, a quantity that can be obtained as an expectation) which is conjugate to $f_k$ in the MaxEnt problem.

In the following, we present an expression which will allow us to evaluate $\vec{\lambda}$ in a more practical way.

## III. AN IDENTITY CONCERNING THE LAGRANGE MULTIPLIERS

In order to obtain a direct expression for $\vec{\lambda}$ we proceed in the following manner. First, we imagine a surface $\Sigma$ defined by the condition $B(\vec{x}) = B_0$, with $B$ a differentiable function of $N$ variables. The surface $\Sigma$ encloses a volume $V$, such that for points in $V$, $B(\vec{x}) < B_0$. Now consider the expectation value of a quantity (until now arbitrary) $A(\vec{x})$ inside $V$, where the expectation is taken using the updated distribution $P(\vec{x})$,

$$\langle A(\vec{x})\rangle_{\vec{\lambda},V} = \frac{1}{\mathcal{Z}}\int_V d\vec{x}\pi(\vec{x})e^{-\vec{\lambda}\cdot\vec{f}(\vec{x})}A(\vec{x}). \tag{7}$$

Suppose $A$ is the divergence of a vector field $\vec{v}$,

$$A(\vec{x}) = \nabla\cdot\vec{v}(\vec{x}), \tag{8}$$

and denote

$$u(\vec{x}) = \pi(\vec{x})e^{-\vec{\lambda}\cdot\vec{f}(\vec{x})}. \tag{9}$$

We can apply the divergence theorem,

$$\int_V d\vec{x}u\nabla\cdot\vec{v} = \int_\Sigma d\Sigma\hat{n}\cdot u\vec{v} - \int_V d\vec{x}\vec{v}\cdot\nabla u, \tag{10}$$

where, in this case, $\hat{n} = \frac{\nabla B}{|\nabla B|}$. This leads to

$$\langle\nabla\cdot\vec{v}\rangle_{\vec{\lambda},V} = \frac{1}{\mathcal{Z}}\int_\Sigma d\Sigma e^{-\vec{\lambda}\cdot\vec{f}}\left(\frac{\vec{v}\cdot\nabla B}{|\nabla B|}\right) + \frac{1}{\mathcal{Z}}\int_V d\vec{x}e^{-\vec{\lambda}\cdot\vec{f}}[\pi(\vec{x})(\mathbb{J}^T\vec{\lambda})\cdot\vec{v} - \nabla\pi(\vec{x})\cdot\vec{v}],$$

where $\mathbb{J}$ denotes the Jacobian matrix, $J_{i,j} = \partial f_i/\partial x_j$. The surface integral on the right-hand side can be expressed in terms of a Dirac $\delta$ function,

$$\int_\Sigma d\Sigma e^{-\vec{\lambda}\cdot\vec{f}}\left(\frac{\vec{v}\cdot\nabla B}{|\nabla B|}\right) = \int d\vec{x}\pi(\vec{x})e^{-\vec{\lambda}\cdot\vec{f}}\delta(B_0 - B(\vec{x}))\vec{v}\cdot\nabla B. \tag{11}$$

Using that

$$\delta(B_0 - B(\vec{x})) = \frac{\partial}{\partial B_0}\Theta(B_0 - B(\vec{x})),$$

we finally arrive at

$$\langle\nabla\cdot\vec{v} - (\mathbb{J}^T\vec{\lambda})\cdot\vec{v} + \vec{v}\cdot\nabla\ln\pi(\vec{x})\rangle_{\vec{\lambda},V} = \frac{\partial}{\partial B_0}\langle\vec{v}\cdot\nabla B\rangle_{\vec{\lambda},V}, \tag{12}$$

where the left-hand size includes $\vec{\lambda}$ explicitly. This last result, Eq. (12), should be valid as long as $B$ is differentiable and $\mathbb{J}$ exists. In particular, if we take $V$ to represent the whole volume (i.e., we take $B_0$ large enough so $B(\vec{x}) < B_0$ for all $\vec{x}$), then the right-hand side vanishes, and we obtain

$$\langle\nabla\cdot\vec{v}\rangle_{\vec{\lambda}} = \langle(\mathbb{J}^T\vec{\lambda})\cdot\vec{v} - \vec{v}\cdot\nabla\ln\pi(\vec{x})\rangle_{\vec{\lambda}}. \tag{13}$$

Equations (12) and (13) are the main results of this paper, and they will, in the following, be jointly referred to as the conjugate variable theorem (CVT). The CVT connects expectation values involving the prior probability $\pi(\vec{x})$, the Lagrange multipliers $\vec{\lambda}$, and the Jacobian of the functions used as constraints in the MaxEnt problem by means of an arbitrary "trial" vector field $\vec{v}$. It can be seen that Eqs. (12) and (13) are *linear* in $\lambda$, so with a proper choice of $m$ different functions $\vec{v}_i(\vec{x})$ the problem of determination of Lagrange multipliers can be written as the following $m\times m$ linear system,

$$\langle\nabla\cdot\vec{v}_i + \vec{v}_i\cdot\nabla\ln\pi\rangle = \sum_{j=1}^m\lambda_j\langle\vec{v}_i\cdot\nabla f_j\rangle, \tag{14}$$

where, for simplicity, the equivalence

$$\mathbb{J}^T\vec{\lambda} = \sum_{k=1}^m\lambda_k\nabla f_k \tag{15}$$

was used.

In practice, in order to be able to solve this system, we need to compute the matrix elements $A_{ij} = \langle \vec{v}_i \cdot \nabla f_j \rangle$ and the elements of the column vector $b_i = \langle \nabla \cdot \vec{v}_i + \vec{v}_i \cdot \nabla \ln \pi \rangle$, which are implicitly dependent on $\vec{\lambda}$ through the probability distribution itself. The success of Eq. (14) as a method for the determination of $\vec{\lambda}$ relies on being able to compute any expectation value (which is possible if we have a large enough sample of the microstates $\{\vec{x}\}$ for averages to converge to expectation values, for instance, from molecular dynamics or Monte Carlo data) or, alternatively, to adequately choosing the trial functions $\vec{v}_i$ so $A_{ij}$ and $b_i$ reduce to either constants or known functions of the $F_i$. This means Eq. (14) does not constitute a universally valid replacement for Eq. (4). It is still applicable, however, in other cases, particularly as an identity relating expectation values and for constructing estimators of $\vec{\lambda}$, as we will describe in the following sections.

The theorem just proved [Eq. (12)] resembles the well-known fluctuation-dissipation theorem in thermodynamics [11,12],

$$\langle (\mathcal{H} - E)^2 \rangle = \frac{\partial E}{\partial \beta}, \tag{16}$$

where $E = \langle \mathcal{H} \rangle$. This relation connects the equilibrium fluctuations of energy in a system and the nonequilibrium linear response relaxation of the system. However, as shown by Jaynes [2], the theorem is simply a consequence of MaxEnt and, therefore, is valid outside thermodynamics in the form

$$\langle g f_k \rangle - G F_k = -\frac{\partial G}{\partial \lambda_k}, \tag{17}$$

where $g(\vec{x})$ is an arbitrary function and $G = \langle g \rangle$. For a single constraint and setting $g = f$, we get the usual form,

$$\langle (f - F)^2 \rangle = -\frac{\partial F}{\partial \lambda}, \tag{18}$$

from which follows Eq. (16) as a particular case. The fluctuation–dissipation theorem can be combined with the CVT as shown in the third example of Sec. V.

It is interesting to note that the condition of maximum entropy subjected to constraints is mathematically equivalent to the condition of validity of the CVT (i.e., CVT holds if and only if the distribution $P$ has the MaxEnt form). To prove this, consider Eq. (13) in the context of an arbitrary probability distribution $\tilde{P}(\vec{x})$, not assumed to be of the MaxEnt form. Apply the substitution $\vec{v} = \vec{\omega}\delta(\vec{x} - \vec{x}_0)$ with $\vec{\omega}$ an arbitrary constant vector. We have

$$\vec{\omega} \cdot \langle \nabla \delta(\vec{x} - \vec{x}_0) \rangle = \vec{\omega} \cdot \left\langle \delta(\vec{x} - \vec{x}_0) \left( \sum_k \lambda_k \nabla f_k - \nabla \ln \pi \right) \right\rangle, \tag{19}$$

where now the expectation is taken over the arbitrary probability distribution $\tilde{P}$.

As Eq. (19) must be valid for any choice of $\vec{\omega}$, we can drop out $\vec{\omega}$ from both sides of the equation and, using the identities,

$$\langle g(\vec{x}) \delta(\vec{x} - \vec{x}_0) \rangle = g(\vec{x}_0) \tilde{P}(\vec{x}_0), \tag{20}$$

for $g$ an arbitrary function of $\vec{x}$ and

$$\langle \nabla \delta(\vec{x} - \vec{x}_0) \rangle = -\nabla \tilde{P}(\vec{x}_0) \tag{21}$$

we obtain

$$\frac{\nabla \tilde{P}(\vec{x}_0)}{\tilde{P}(\vec{x}_0)} = -\sum_k \lambda_k \nabla f_k(\vec{x}_0) + \nabla \ln \pi(\vec{x}_0). \tag{22}$$

This last expression can be written as

$$\nabla \left( \ln \tilde{P}(\vec{x}_0) + \sum_k \lambda_k f_k(\vec{x}_0) - \ln \pi(\vec{x}_0) \right) = 0, \tag{23}$$

from which it follows that

$$\tilde{P}(\vec{x}_0) \propto \pi(\vec{x}_0) e^{-\sum_k \lambda_k f_k(\vec{x}_0)}. \tag{24}$$

In other words, $\tilde{P}$ must be a MaxEnt solution (or, in Fisher statistics, $\tilde{P}$ must have sufficient statistics, see the Pitman-Koopman-Darmois theorem [13–15]). Therefore, we have proved that Eqs. (2) and (13) are equivalent, in the sense that one implies the other. The same kind of equivalence can be proven in a straightforward way between the fluctuation-dissipation theorem [Eq. (17)] and Eq. (2) by using $g = \delta(\vec{x} - \vec{x}_0)$.

## IV. APPLICATIONS

From the above Eqs. (12) and (13) it is possible to derive several particular cases known in statistical mechanics, as well as new relations, as far as we know, not previously reported and applicable to other problems of maximum entropy outside physics. Here we explore the case of a single constraint and its particular cases.

For the case of $\vec{f}(\vec{x})$ being a simple scalar $f(\vec{x})$, the Lagrange multiplier $\lambda$ is given directly by using Eq. (13) [or, equivalently, Eq. (14)] with $m = 1$,

$$\lambda = \frac{\langle \nabla \cdot \vec{v} + \vec{v} \cdot \nabla \ln \pi \rangle_\lambda}{\langle \vec{v} \cdot \nabla f \rangle_\lambda}. \tag{25}$$

Furthermore, if the ignorance prior $\pi$ is flat, i.e., $|\nabla \pi| = 0$, then

$$\lambda = \frac{\langle \nabla \cdot \vec{v} \rangle_\lambda}{\langle \vec{v} \cdot \nabla f \rangle_\lambda}. \tag{26}$$

Choosing $\vec{v} = \hat{e}_k x_j$ and $f = \mathcal{H}$, this reduces to the so-called *hypervirial* relations in statistical mechanics [5,16],

$$\delta_{kj} = \beta \left\langle x_j \frac{\partial \mathcal{H}}{\partial x_k} \right\rangle. \tag{27}$$

In this context it is clear that this formula, related to the equipartition of energy, is only one of many possible estimators for the inverse temperature $\beta$, which is, of course, the Lagrange multiplier in the canonical distribution,

$$P(\vec{\Gamma}) = \frac{1}{\mathcal{Z}(\beta)} e^{-\beta \mathcal{H}(\vec{\Gamma})}. \tag{28}$$

On the other hand, if we choose $\vec{v} = \vec{\omega}/(\vec{\omega} \cdot \nabla f)$ with $\vec{\omega}$, an arbitrary vector such that $\vec{\omega} \cdot \nabla f$ never vanishes [Eq. (26)] reduces to

$$\lambda = \left\langle \nabla \cdot \frac{\vec{\omega}}{\vec{\omega} \cdot \nabla f} \right\rangle_\lambda. \tag{29}$$

In the context of microcanonical statistical mechanics [17], $\langle \mathcal{H} \rangle = E$ is given as a constraint, that is, $f = \mathcal{H}$. Choosing

$\vec{\omega} = \nabla\mathcal{H}$, we obtain

$$\beta = \left\langle \nabla \cdot \frac{\nabla\mathcal{H}}{|\nabla\mathcal{H}|^2} \right\rangle_E, \tag{30}$$

which is Rugh's expression for the dynamical temperature [6]. This new observable, constructed from the derivatives of the Hamiltonian, is widely used in computer simulation, either for monitoring the instantaneous temperature and the approach to equilibrium in a molecular simulation [18] or as a thermostat allowing the implementation of molecular dynamics in the canonical ensemble [19]. It has also led to the insight that spin degrees of freedom [20], whose Hamiltonian lacks a kinetic energy term, can have a dynamical temperature associated to them as well. Microcanonical temperature has also been used in the context of Yang-Mills models [21] and Bose systems [22]. For a complete review of the importance of dynamical temperature, see Ref. [23].

In general, from Eq. (29) we can define a family of estimators

$$\hat{\lambda} = \nabla \cdot \frac{\vec{\omega}}{\vec{\omega} \cdot \nabla f} \tag{31}$$

such that its expectation value $\langle\hat{\lambda}\rangle$ corresponds to the value of the Lagrange multiplier $\lambda$. Therefore, Eq. (30) is only one such estimator of the inverse temperature. In fact, Rickayzen [7] found Eq. (31) from particular properties of the microcanonical ensemble.

In a Bayesian formulation, the fact that the observable $\lambda$ has a family of definitions given by Eq. (31) is reasonable and far from being controversial. If, however, we assume that there is an intrinsic property $\hat{\lambda}(\vec{x})$ of the microstate, independent of the observer doing the inference, we should expect the function $\hat{\lambda}$ to be unique, and then Rickayzen's expression [our Eq. (31)] for the inverse temperature is difficult to reconcile with this assumption.

Let us make a general point about the estimators of $\lambda$ given by Eq. (31). Any estimator $\hat{\lambda}$ from this family serves as a conjugate *function* to $f(\vec{x})$, in the same sense that $\lambda$ is a conjugate *quantity* to $F$. Notice that, in one dimension ($\vec{x} \to x$) there is only one possible estimator [as $\omega(x)$ cancels out of Eq. (31)],

$$\hat{\lambda}(x) = -\frac{f''(x)}{(f'(x))^2}. \tag{32}$$

This is also valid if in Eq. (31) we take $\vec{\omega}$, having a single nonzero constant component $\omega_k$, the estimator then reduces to

$$\hat{\lambda}_k(\vec{x}) = -\frac{\partial^2 f/\partial x_k^2}{(\partial f/\partial x_k)^2}. \tag{33}$$

This fact hints towards a geometrical interpretation of the conjugate $\hat{\lambda}$ as some kind of curvature of $f(\vec{x})$ along the particular direction $\vec{\omega}$. Notice also that the magnitude of $\lambda$ (seen as a Lagrange multiplier in an optimization problem) is related to the "strength" of the constraint $\langle f\rangle = F$. In information-theoretical terms, the more informative the constraint is, the larger the absolute value of $\lambda$.

A practical question is how these different estimators $\hat{\lambda}$ compare to each other in terms of accuracy. In order to evaluate this, we can use the concept of statistical efficiency $e[\hat{\lambda}]$, given by [24]

$$e[\hat{\lambda}] = \frac{1}{I_F(\lambda)\sigma^2[\hat{\lambda}]} \leqslant 1, \tag{34}$$

where $\sigma^2[\hat{\lambda}] = \langle(\hat{\lambda} - \lambda)^2\rangle$ is the variance of the estimator, $I_F(\lambda)$ is the Fisher information, given by

$$I_F(\lambda) = \left\langle \left( \frac{\partial}{\partial\lambda} \ln P(\vec{x}) \right)^2 \right\rangle = \langle (f - F)^2 \rangle, \tag{35}$$

and the inequality is known as the Crámer-Rao bound. In physicists' terms, an estimator is more efficient (accurate) if it has fewer fluctuations around its average value, and the smallest value of the fluctuations corresponds to the Fisher information. This optimum is attained by the maximum likelihood estimators.

In our context, by using $\vec{v} = g\nabla f/|\nabla f|^2$ in Eq. (26), $g$ an arbitrary function of $\vec{x}$, we arrive at

$$\lambda\langle g\rangle - \langle\hat{\lambda} g\rangle = \left\langle \frac{\nabla g \cdot \nabla f}{|\nabla f|^2} \right\rangle. \tag{36}$$

Choosing $g = \hat{\lambda}$, we get an expression for the variance of the estimator,

$$\sigma^2[\hat{\lambda}] = -\left\langle \frac{\nabla\hat{\lambda} \cdot \nabla f}{|\nabla f|^2} \right\rangle. \tag{37}$$

This allows the determination of the efficiency of the estimator associated with any particular substitution of $\vec{\omega}$. From the upper bound of Eqs. (34) and (37) it follows that the most efficient choice of $\vec{\omega}$ (i.e., the one leading to the maximum likelihood estimator) should satisfy

$$\langle (f - F)^2 \rangle \left\langle \frac{\nabla\hat{\lambda}[\vec{\omega}] \cdot \nabla f}{|\nabla f|^2} \right\rangle = -1. \tag{38}$$

Interestingly, if the estimator $\hat{\lambda}$ depends on $\vec{x}$ only through $f$ itself, i.e., $\hat{\lambda}(\vec{x}) = \hat{\lambda}(f(\vec{x}))$, we then obtain a "fluctuation–dissipation" relation for $\hat{\lambda}$,

$$\langle(\hat{\lambda} - \lambda)^2\rangle = -\left\langle \frac{\partial\hat{\lambda}}{\partial f} \right\rangle, \tag{39}$$

analogous to Eq. (18).

## V. SOME ILLUSTRATIONS OF THE USE OF THE CVT

### A. Recovering the normal distribution

Consider a continuous variable $x$ with known mean $A$ and variance $B$. The MaxEnt distribution $P(x|A,B)$ should have the form

$$P(x|\lambda_A,\lambda_B) = \frac{1}{Z(\lambda_A,\lambda_B)} e^{-\lambda_A x - \lambda_B x^2}. \tag{40}$$

We wish to determine $\lambda_A$ and $\lambda_B$ in terms of $A$ and $B$, without employing Eq. (4) or the maximum likelihood method.

We can recover the usual estimators for $\lambda_A$ and $\lambda_B$ just from the CVT, without solving an optimization problem, as follows. The constraints imposed on the mean and variance lead to the functions $f_A = x$ and $f_B = x^2$. We can write two different

versions of the CVT, using trial functions $v_1$ and $v_2$,

$$\langle v_1'(x)\rangle = \lambda_A\langle v_1(x)f_A'(x)\rangle + \lambda_B\langle v_1(x)f_B'(x)\rangle, \quad (41)$$

$$\langle v_2'(x)\rangle = \lambda_A\langle v_2(x)f_A'(x)\rangle + \lambda_B\langle v_2(x)f_B'(x)\rangle, \quad (42)$$

obtaining a system of linear equations that can be solved for $\lambda_A$ and $\lambda_B$ in terms of some particular expectation values. We expect that some choice will put everything in terms of known quantities such as $A$, $B$ and constants. Let us choose $v_1(x) = 1$ and $v_2(x) = x$; we then get

$$0 = \lambda_A + 2\lambda_B A, \quad (43)$$

$$1 = \lambda_A A + 2\lambda_B B. \quad (44)$$

Solving for $\lambda_A$ and $\lambda_B$, we obtain

$$\lambda_A = -\frac{A}{B - A^2}, \quad (45)$$

$$\lambda_B = \frac{1}{2(B - A^2)}, \quad (46)$$

which are precisely the maximum likelihood estimates. This can be seen by calling $B - A^2 = \langle x^2\rangle - \langle x\rangle^2 = S^2$ and $A = \langle x\rangle = X$ and replacing the estimates for $\lambda_A$ and $\lambda_B$ in the MaxEnt solution [Eq. (40)],

$$P(x|X,S) = \frac{1}{Z(X,S)} e^{\frac{X}{S^2}x - \frac{1}{2S^2}x^2}. \quad (47)$$

Completing the square in the exponential, we get precisely the normal distribution (albeit without the normalization constant, which can be obtained as usual by integration),

$$P(x|X,S) = \frac{1}{\eta(X,S)} e^{-\frac{1}{2S^2}(x-X)^2}, \quad (48)$$

where $\eta(X,S) = Z(X,S)e^{-\frac{X^2}{2S^2}}$.

### B. A single power-law constraint

Consider a continuous variable $x$ with known expectation value of its $n$-th power ($n$ not necessarily an integer), $\langle x^n\rangle = F$. The MaxEnt solution is

$$P(x|\lambda) = \frac{1}{Z(\lambda)} e^{-\lambda x^n}, \quad (49)$$

and we wish to determine $\lambda$ as a function of $F$. For this, we use the CVT with a single trial function $v(x)$,

$$\langle v'(x)\rangle = \lambda\langle v(x)f'(x)\rangle = \lambda n\langle v(x)x^{n-1}\rangle. \quad (50)$$

Using $v(x) = x$, we get

$$\lambda = \frac{1}{nF}. \quad (51)$$

To check this result in an alternative way, we can explicitly compute the partition function

$$Z(\lambda) = \int_0^\infty dx e^{-\lambda x^n} = \lambda^{-1/n}\Gamma(1 + 1/n) \quad (52)$$

and apply Eq. (4),

$$-\frac{\partial}{\partial\lambda}\ln(\lambda^{-1/n}) = \frac{1}{n\lambda} = F, \quad (53)$$

which leads, again, to Eq. (51).

### C. Drawing spheres from two boxes

Consider the following statistical problem. We have two boxes, each filled with a mixture of spheres of different volumes. For the first box, we know the average volume of a sphere is $V_1$ and the average radius is $R_1$. If for the second box we happen to know the average radius $R_2$, can we estimate the average volume $V_2$?

We model the distribution $P(r)$ of either box as a MaxEnt distribution with known $\langle r^3\rangle$,

$$P(r|\lambda) = \frac{1}{Z(\lambda)} e^{-\lambda r^3}\Theta(r), \quad (54)$$

with the parameter $\lambda$ completely characterizing a box. The motivation for this model could be that the volume of a particular sphere $V(r) = (4\pi/3)r^3$ should completely characterize the chances of it being picked out of the box.

Use of the CVT gives

$$\langle v'(r)\rangle = 3\lambda\langle v(r)r^2\rangle - \frac{v(0)}{Z(\lambda)}, \quad (55)$$

and combining the results from using $v(r) = r$ and $v(r) = r^2$, we get

$$\lambda = \frac{1}{3\langle r^3\rangle}, \quad (56)$$

$$2\langle r\rangle = 3\lambda\langle r^4\rangle. \quad (57)$$

From the fluctuation-dissipation theorem [Eq. (18)],

$$\langle r^4\rangle - \langle r\rangle\langle r^3\rangle = -\frac{\partial}{\partial\lambda}\langle r\rangle, \quad (58)$$

and then, replacing $\langle r^4\rangle$ and $\langle r^3\rangle$, we obtain a differential equation for $\langle r\rangle$,

$$\langle r\rangle = -3\lambda\frac{\partial}{\partial\lambda}\langle r\rangle. \quad (59)$$

The solution is

$$\langle r\rangle = A\lambda^{-1/3} = A(3\langle r^3\rangle)^{1/3}, \quad (60)$$

where $A$ is an integration constant, independent of $\lambda$. The information we have on the first box, namely $\langle r^3\rangle_1 = (3V_1/4\pi)$ and $\langle r\rangle_1 = R_1$, fixes $A$ to be

$$A = \frac{R_1}{(9V_1/4\pi)^{1/3}}, \quad (61)$$

and then, given $R_2 = \langle r\rangle_2$ for the second box, we get its expected volume,

$$V_2 = (4\pi/3)\langle r^3\rangle_2 = V_1\left(\frac{R_2}{R_1}\right)^3. \quad (62)$$

This solves our question in the affirmative. However, Eq. (62), as every MaxEnt prediction, is not a logical deduction but a plausible inference and, therefore, not guaranteed to be true beforehand. If we replicate this experiment (in real life or through a computer simulation) and find that $A$ differs for two boxes, and this means we have learned something: We did not include all the information needed to describe the contents of the box (in more precise terms, $r^3$ is not a sufficient statistic for the problem). In other words, it would mean the assumption of Eq. (54) is at fault; that is, the chances of drawing a sphere

depend not only on the volume of the sphere but on other factors, such as the remaining contents of the box.

If we are interested in the exact value of the constant $A$, we can compute $\langle r\rangle$ explicitly as a function of $\lambda$, obtaining

$$A = \lambda^{1/3}\frac{\int_0^\infty dr\cdot re^{-\lambda r^3}}{\int_0^\infty dre^{-\lambda r^3}} = \frac{\Gamma(2/3)}{3\Gamma(4/3)}. \tag{63}$$

This, however, was not needed for our problem, as $A$ was "measured empirically" using the first box. It is interesting to note that

$$\langle V\rangle = \left(\frac{1}{3A^3}\right)\frac{4\pi}{3}\langle r\rangle^3, \tag{64}$$

which means that, contrary to a naïve intuition, the expected volume is larger than the volume of a sphere with the expected radius by a factor $(1/3A^3)$, approximately 2.58. This is because the expected volume $\langle V\rangle$ is more sensitive to drawing a sphere with large radius (due to the cubic dependence) than the expected radius $\langle r\rangle$.

## VI. CONCLUSIONS

We have derived a general theorem, Eq. (12), connecting the values of Lagrange multipliers in MaxEnt inference to expectation values related to an arbitrary trial function. This theorem provides for some particular cases an alternative "shortcut" to the use of Eq. (4) involving the logarithmic derivatives of the partition function, which leads to a system of nonlinear equations. For these cases, our result provides the corresponding linear system, which could be applicable in many problems when the appropriate expectations are known. Equation (12), which, in particular cases, reduces to the hypervirial relations of statistical mechanics and to the expression for the temperature in the microcanonical ensemble obtained by Rugh and generalized by Rickayzen, is just a consequence of the maximum entropy formalism for probabilistic reasoning. Therefore, its validity is independent of ergodicity considerations or equal *a priori* probabilities for the microstates in a thermodynamic system, even of the existence of such thermodynamic system: It is valid whenever we have reasoning under incomplete information provided as expectation values. We show also that this equation is equivalent to the maximum entropy condition itself (or, in Fisher statistics terminology, to the existence of a sufficient statistic). Therefore, from the point of view of frequentist statistics, Eq. (12) can become an alternative tool for the estimation of parameters in any probability distribution with a sufficient statistic. Three examples in statistics illustrate the use of the CVT in different scenarios.

## ACKNOWLEDGMENTS

S.D. acknowledges support from Fondecyt 3110017. We gratefully appreciate the suggestions of two anonymous referees.

[1] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[2] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, UK, 2003).
[3] A. Caticha and A. Giffin, AIP Conf. Proc. **872**, 31 (2006).
[4] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, Phys. Rev. A **30**, 2638 (1984).
[5] C. G. Gray and K. E. Gubbins, *Theory of Molecular Fluids: Fundamentals* (Oxford University Press, Oxford, UK, 1984).
[6] H. H. Rugh, Phys. Rev. Lett. **78**, 772 (1997).
[7] G. Rickayzen and J. G. Powles, J. Chem. Phys. **114**, 4333 (2001).
[8] A. Baranyai, J. Chem. Phys. **112**, 3964 (2000).
[9] N. Rathore, T. A. Knotts, and J. J. de Pablo, Biophys. J. **85**, 3963 (2003).
[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2006).
[11] H. Nyquist, Phys. Rev. **32**, 110 (1928).
[12] H. B. Callen and T. A. Welton, Phys. Rev. **83**, 34 (1951).
[13] E. J. G. Pitman, Proc. Cambridge Philos. Soc. **32**, 567 (1936).
[14] B. O. Koopman, Trans. Am. Math. Soc. **39**, 399 (1936).
[15] G. Darmois, Rev. Inst. Int. Stat. **13**, 9 (1945).
[16] K. Huang, *Statistical Mechanics* (Wiley, New York, 1987).
[17] The microcanonical ensemble can be obtained from MaxEnt under known mean and variance of the energy, which leads to a Gaussian distribution, then taking the limit of zero variance to recover the usual $P(\vec{x}) = \delta(\mathcal{H} - E)/\Omega(E)$.
[18] B. D. Butler, G. Ayton, O. G. Jepps, and D. J. Evans, J. Chem. Phys. **109**, 6519 (1998).
[19] C. Braga and K. P. Travis, J. Chem. Phys. **123**, 134101 (2005).
[20] W. B. Nurdin and K. D. Schotte, Phys. Rev. E **61**, 3579 (2000).
[21] V. M. Bannur, Phys. Rev. C **72**, 024904 (2005).
[22] P. B. Blakie and M. J. Davis, J. Phys. B **40**, 2043 (2007).
[23] J. G. Powles, G. Rickayzen, and D. M. Heyes, Mol. Phys. **103**, 1361 (2005).
[24] R. W. Keener, *Statistical Theory: Notes for a Course in Theoretical Statistics* (Springer, Berlin, 2006).