Eur. Phys. J. B (2016) 89: 124
DOI: 10.1140/epjb/e2016-70175-6

THE EUROPEAN PHYSICAL JOURNAL B

Topical Review

Maximum information entropy principle and the interpretation of probabilities in statistical mechanics — a short review

Domagoj Kuić[a]

University of Split, Faculty of Science, R. Boškovića 33, 21000 Split, Croatia

Received 19 March 2016 / Received in final form 29 March 2016
Published online 16 May 2016 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2016

Abstract. In this paper an alternative approach to statistical mechanics based on the maximum information entropy principle (MaxEnt) is examined, specifically its close relation with the Gibbs method of ensembles. It is shown that the MaxEnt formalism is the logical extension of the Gibbs formalism of equilibrium statistical mechanics that is entirely independent of the frequentist interpretation of probabilities only as factual (i.e. experimentally verifiable) properties of the real world. Furthermore, we show that, consistently with the law of large numbers, the relative frequencies of the ensemble of systems prepared under identical conditions (i.e. identical constraints) actually correspond to the MaxEnt probabilites in the limit of a large number of systems in the ensemble. This result implies that the probabilities in statistical mechanics can be interpreted, independently of the frequency interpretation, on the basis of the maximum information entropy principle.

1 Introduction

From the point of view of predictive statistical mechanics which is based on the maximum information entropy principle, with the exception of quantum mechanical probabilities, there is no reason to consider some particular probability distribution as the true distribution describing the system [1]. Such a view is in a marked contrast to the interpretation that defines probability only in terms of the limit of a relative frequency of the outcome in an infinite sampling sequence, where the probabilities are therefore factual properties of the observed system [2]. From the law of large numbers it follows that the relative frequency of success in a sequence of e.g. Bernoulli trials (in a sequence of repeated independent trials of an experiment with only two possible outcomes) converges to the theoretical probability. For example, a fair coin toss is a Bernoulli trial where the theoretical probability that the outcome will be heads is equal to 1/2. According to the law of large numbers, the proportion of heads in a large number of fair coin tosses will converge to 1/2 as the number of tosses approaches infinity. This means convergence in probability in the weak form of the law and convergence with probability one in the strong form, where the strong form of the law always implies the weak form of the law [3]. Accordingly, the relative frequency is a factual property of the real world that can be measured by repeating a large number of trials, or estimated from the theoretical probability. Probability, on the other hand, is something that we assign to individual events, or we calculate it for the composite events according to the rules (axioms) of probability theory, from the previously assigned probabilities of individual events.

In different applications of statistical mechanics, we try to predict the results of, or draw inferences from, some experiment that can be repeated indefinitely under what appears to be identical conditions (i.e. on the ensemble of identically prepared systems). Although traditional expositions of statistical mechanics such as [4] define the probability as the limiting relative frequency in independent repetitions of a statistical experiment, the relation between frequencies and probabilities, implied by the law of large numbers, in statistical mechanics becomes very complex, because in reality for a macroscopic system, we do not measure the relative frequency of the occurrence of its individual microscopic states in a sequence of infinite or a large number of trials.

In the frequentist interpretation probabilities are always experimentaly verifiable, and consequently, one of the foundational problems of statistical mechanics would be to derive and to justify the probabilities of microscopic events, in the sense of frequencies in the ensemble of indentically prepared systems, from the first principles i.e. from equations of motion. This is the main problem of ergodic theory approach to statistical mechanics [5,6]. Jaynes presented the opposite view, that if we choose to represent only the degree of our knowledge about the individual system, then there can not be anything physically real in the frequencies in the corresponding ensemble of a large number of systems, nor there is any sense in asking which ensemble is the only correct one [7]. In the interpretation given by Jaynes, what we call different ensembles

[a] e-mail: dkuic@pmfst.hr

corresponds in reality to different degrees of knowledge about the individual system, or about some physical situation. In the argumentation of this viewpoint, Jaynes referred to the statement by Gibbs, according to which the ensembles are chosen only to illustrate the probabilites of events in the real word [7,8].

The simplest interpretation of the Gibbs method of ensembles and the MaxEnt formalism follows from the fact that by maximizing the information entropy, which is also known as the uncertainty represented by a probability distribution, subject to given macroscopic constraints, one predicts just the macroscopic behaviour that can happen in the greatest number of microscopic realizations (i.e. greatest multiplicity) compatible with those constraints [7,9–11]. Without going deeper into the problem of interpretation of probabilites, which is even more pronounced in the case of nonequilibrium states, it is more important that the distributions obtained from the application of the principle of maximum information entropy depend only on the available information and do not depend on arbitrary assumptions related to missing information. If we refer only to predictions, from the same viewpoint one can speak about the objectivity only in the extent in which the incompleteness of information about the system is taken into account. Consistent with this way of thinking, by applying the principle of maximum information entropy, we come to the relevant statistical distributions, and this is the subject of the paper.

The structure of the paper is as follows. Section 2 is a brief introduction on the Shannon's concept of information entropy [12], and on the principle of maximum information entropy and MaxEnt formalism formulated by Jaynes [13,14]. Section 3 deals with the interpretation of MaxEnt formalism in statistical mechanics as given by Jaynes [13] and Grandy [15,16]. Section 4 introduces the independent interpretation of probabilities in statistical mechanics on the basis of the principle of maximum information entropy. We modify here and extend the analysis given by Jaynes in reference [17] and show that it has important consequences for the interpretation of probabilities. Section 5 is the conclusion summarizing the main results of the paper.

2 Information entropy — measure of uncertainty — and the principle of maximum information entropy

In Shannon's information theory [12] the quantity of the form

$$H(p_1,\ldots,p_n) = -K\sum_{i=1}^{n} p_i \log p_i, \tag{1}$$

has a central role as a measure of information, choice and uncertainty for different probability distributions $p_1,\ldots,p_n$. Starting from the understanding that the problem of constructing a communication device depends on the statistical structure of the information that is to be communicated (i.e. on the probabilities $p_1,p_2,\ldots,p_n$ of the symbols $A_1,A_2,\ldots,A_n$ of some alphabet) Shannon gave until that time the most general definition of the measure of amount of information. Sequences of symbols or "letters" may form the set of "words" of certain length, and the amount of information is measured analogously. Positive constant K in equation (1) depends on the choice of a unit for the amount of information. In real applications expression (1), with the logarithmic base 2 and $K = 1$, represents the expected number of bits per symbol necessary to encode the random signal forming a memoryless source. But most importantly, Shannon's interpretation of the function (1) is not dependent on the specific context of information theory. He defined the function (1) as the measure of *uncertainty* related to the occurrence of possible events, or more specifically, as a measure of uncertainty *represented by the probability distribution* $p_1,p_2,\ldots,p_n$. This is substantiated by three reasonable properties that are required from such a measure $H(p_1,\ldots,p_n)$ that are sufficient to uniquely determine the form of this function: continuity, monotonic increase with number of possibilities in case when all probabilities are equal, and the unique and consistent composition law for the addition of uncertainties when mutually exclusive events are grouped into composite events. Shannon called the function (1) the entropy of the set of probabilities $p_1,p_2,\ldots,p_n$.

However, we have still not answered an open question on how to determine or to choose the appropriate probability distribution for a particular problem or a system. The principle of maximum information entropy (MaxEnt) was formulated by Jaynes [13,14] as a general criterion for construction of the probability distribution when the available information is not sufficient for the unique determination of the distribution. This principle is based on the following rationale: maximization of the information entropy (the uncertainty) subject to given constraints includes in the probability distribution only the information represented by these constraints. Therefore, predictions derived from such a probability distribution depend only on the available information and do not depend on arbitrary assumptions related to missing information.

The mathematical formulation of this principle is known as the MaxEnt algorithm. Let us consider it on the following example. Let the variable x takes n values $\{x_1,\ldots,x_n\}$ with probabilities $\{p_1,\ldots,p_n\}$ and the only data available are given by the expectation values of the functions $f_k(x)$:

$$F_k = \langle f_k(x)\rangle = \sum_{i=1}^{n} p_i f_k(x_i), \quad k = 1,2,\ldots,m < n. \tag{2}$$

Probability distribution must also satisfy the normalization condition

$$\sum_{i=1}^{n} p_i = 1, \qquad (p_i \geq 0, \quad i = 1,2,\ldots,n). \tag{3}$$

In most cases the available information given by the set of equations (2) is far less then sufficient for the unique determination of the set of probabilities $\{p_1,\ldots,p_n\}$, i.e. $m \ll n-1$. In such cases, probability distribution $\{p_1,\ldots,p_n\}$

is determined by applying the MaxEnt principle. Probability distribution $\{p_1, \ldots, p_n\}$ for which the information entropy (1) is maximum subject to the constraints (2) is found by the method of Lagrange multipliers, i.e. by maximizing the function

$$I = -\sum_{i=1}^{n} p_i \log p_i - (\lambda_0 - 1)\left(\sum_{i=1}^{n} p_i - 1\right) - \sum_{k=1}^{m} \lambda_k \left(\sum_{i=1}^{n} p_i f_k(x_i) - F_k\right), \quad (4)$$

where $\lambda_0 - 1$, λ_k, $k = 1, 2, \ldots, m$, are the Lagrange multipliers. In this way we obtain the MaxEnt probability distribution

$$p_i = \frac{1}{Z} \exp\left\{-\sum_{k=1}^{m} \lambda_k f_k(x_i)\right\}, \qquad i = 1, 2, \ldots, n. \quad (5)$$

The normalization factor $Z = e^{\lambda_0}$ which is also known as the partition function is given by:

$$Z \equiv Z(\lambda_1, \ldots, \lambda_m) = \sum_{i=1}^{n} \exp\left\{-\sum_{k=1}^{m} \lambda_k f_k(x_i)\right\}. \quad (6)$$

The expectation values of the functions $\langle f_k(x)\rangle = F_k$, $k = 1, 2, \ldots, m$, given by the conditions (2), are equivalently given also by:

$$F_k = \langle f_k(x)\rangle = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k}, \quad k = 1, 2, \ldots, m. \quad (7)$$

Let us assume that set of $m+1$ equations consisting of m equations (2) and (3) is consistent and that these equations are linearly independent. Then, using (5) and solving this set of equations, one can determine the Lagrange multipliers λ_k, $k = 1, 2, \ldots, m$, as single-valued functions $\lambda_k(F)$ of the expected values $F = (F_1, \ldots, F_m)$. The proof is given in reference [18]. Then, by introducing the MaxEnt probability distribution (5) in the expression (1) for information entropy, the maximum of information entropy subject to the conditions (2) and (3) is obtained as the function of the expected values $F = (F_1, \ldots, F_m)$:

$$(S_I)_{\max} = \log Z(\lambda_1, \ldots, \lambda_m) + \sum_{k=1}^{m} \lambda_k F_k = S(F_1, \ldots, F_m). \quad (8)$$

Assuming that the functions $\lambda_k(F)$, $k = 1, 2, \ldots, m$, are continuously differentiable (or at least piecewise smooth), from equations (7) and (8) it follows that

$$\lambda_k = \frac{\partial S(F_1, \ldots, F_m)}{\partial F_k}, \qquad k = 1, 2, \ldots, m. \quad (9)$$

From equations (7), (8) and (9) it is obvious that the functions $\log Z(\lambda_1, \ldots, \lambda_m)$ and $S(F_1, \ldots, F_m)$ are mutually related by a Legendre transformation. Functions related in this way contain the same information but it is expressed through different variables.

Furthermore, functions $\log Z(\lambda_1, \ldots, \lambda_m)$ and $S(F_1, \ldots, F_m)$ give, in a simple way, the variances and covariances of the functions $f_k(x)$, $k = 1, 2, \ldots, m$. Using (5), (6) and (7) one obtains

$$\frac{\partial^2 \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_l \partial \lambda_k} = -\frac{\partial F_k}{\partial \lambda_l} = -\frac{\partial F_l}{\partial \lambda_k} = \langle f_k(x) f_l(x)\rangle - \langle f_k(x)\rangle\langle f_l(x)\rangle = -A_{kl}, \quad k, l = 1, 2, \ldots, m, \quad (10)$$

where A is a symmetric matrix, $A_{kl} = A_{lk}$. In a similar way, using (9) one obtains

$$\frac{\partial^2 S(F_1, \ldots, F_m)}{\partial F_l \partial F_k} = \frac{\partial \lambda_k}{\partial F_l} = \frac{\partial \lambda_l}{\partial F_k} = B_{kl}, \qquad k, l = 1, 2, \ldots, m, \quad (11)$$

where B is also a symmetric matrix, $B_{kl} = B_{lk}$. Then, from equations (10) and (11) and the chain rule for derivatives, it follows that

$$\frac{\partial \lambda_j}{\partial \lambda_l} = \sum_{k=1}^{m} \frac{\partial \lambda_j}{\partial F_k} \frac{\partial F_k}{\partial \lambda_l} = B_{jk} A_{kl} = \delta_{jl}, \quad j, l = 1, 2, \ldots, m, \quad (12)$$

and similarly,

$$\frac{\partial F_j}{\partial F_l} = \sum_{k=1}^{m} \frac{\partial F_j}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial F_l} = A_{jk} B_{kl} = \delta_{jl}, \quad j, l = 1, 2, \ldots, m. \quad (13)$$

Therefore, the matrices given by equations (10) and (11) are inverses, $A^{-1} = B$.

Elements of the matrix A are the second partial derivatives of the functions $\log Z(\lambda_1, \ldots, \lambda_m)$ and represent the measure of the expected dispersion and mutual correlation of the functions $f_k(x)$, $k = 1, 2, \ldots, m$. Diagonal elements of the matrix A give as the notion about the deviation of the variables $f_k(x)$ from their expectation values $\langle f_k(x)\rangle$. Furthermore, from (5), (6) and (7), it follows that the covariance of some other function $g(x)$ with the function $f_k(x)$ is obtained as:

$$-\frac{\partial \langle g(x)\rangle}{\partial \lambda_k} = \langle g(x) f_k(x)\rangle - \langle g(x)\rangle\langle f_k(x)\rangle, \quad k = 1, 2, \ldots, m. \quad (14)$$

3 Interpretation of MaxEnt formalism in statistical mechanics

It is clear that the MaxEnt probability distribution (5) has the same form as Gibbs ensemble probability distributions from equilibrium statistical mechanics. This is not surprising since the rationale of the Gibbs method of constructing ensembles was to assign that probability distribution which, while agreeing with what is known

(i.e. the data given by constraints), gives the least value of the average index (logarithm) of probability of phase i.e. $\sum_{i=1}^{n} p_i \log p_i$ [7,8]. This procedure has lead Gibbs to the canonical ensemble for closed systems in thermal equilibrium with the environment, the grand canonical ensemble for open systems, and an ensemble for a system rotating at a fixed angular velocity. However, MaxEnt formalism represents a general method of statistical inference which is applicable, in principle, to all problems where only incomplete and partial information about the problem is available. Equations from the last section represent the generic form of the MaxEnt formalism. To give them a physical interpretation they should be put in the context of some specific physical situation. Since the Lagrange multipliers $\lambda = (\lambda_1, \ldots, \lambda_m)$, under certain conditions, are single-valued functions of the expected values $F = (F_1, \ldots, F_m)$, and at the same time the only parameters in the MaxEnt probability distribution, physical interpretation of these quantities is of special pertinence in that sense.

It will now be shown that the physical interpretation of the Lagrange multipliers follows from the relation describing the changes of the expected values. Values of the functions $f_k(x_i)$, $k = 1, 2, \ldots, m$, associated with the values x_i of the variable x, $i = 1, 2, \ldots, n$, can represent the eigenvalues of some specific physical quantities, for example energy eigenvalues E_i, or eigenvalues of the quantities from the set of compatible quantities. Let us assume that the small change in the expectation values $\langle f_k(x)\rangle$ is done by the small change of the functions $f_k(x_i)$ and the probabilities p_i,

$$\delta\langle f_k(x)\rangle = \sum_{i=1}^{n} p_i \delta f_k(x_i) + \sum_{i=1}^{n} f_k(x_i)\delta p_i, \quad k = 1, 2, \ldots, m. \tag{15}$$

Here, $\delta\langle f_k(x)\rangle$ is the change of the expectation value $\langle f_k(x)\rangle$ and $\langle \delta f_k(x)\rangle = \sum_i p_i \delta f_k(x_i)$ is the expectation value of the change of $f_k(x)$. Their difference depends on the changes in the probabilities δp_i,

$$\delta\langle f_k(x)\rangle - \langle \delta f_k(x)\rangle = \sum_{i=1}^{n} f_k(x_i)\delta p_i, \qquad k = 1, 2, \ldots, m. \tag{16}$$

The change of information entropy S_I is equal to

$$\delta S_I = -\sum_{i=1}^{n} \delta p_i \log p_i. \tag{17}$$

Introducing the MaxEnt probabilities (5) for $\{p_i\}$ in equation (17) and using equations (3) and (16), one obtains

$$\begin{aligned} \delta S &= \sum_{k=1}^{m}\sum_{i=1}^{n} \lambda_k f_k(x_i)\delta p_i \\ &= \sum_{k=1}^{m} \lambda_k \left(\delta\langle f_k(x)\rangle - \langle \delta f_k(x)\rangle\right). \end{aligned} \tag{18}$$

Assuming that $\{p_i + \delta p_i\}$ is also a MaxEnt probability distribution, equation (18) then gives the change of the maximum of information entropy due to the change in expected values (i.e. the constraints). The meaning of equation (18) is simple to understand, if we introduce

$$\begin{aligned} \delta Q_k &= \sum_{i=1}^{n} f_k(x_i)\delta p_i = \delta\langle f_k(x)\rangle - \langle \delta f_k(x)\rangle, \\ k &= 1, 2, \ldots, m, \end{aligned} \tag{19}$$

and then using this write δS in the form

$$\delta S = \sum_{k=1}^{m} \lambda_k \delta Q_k. \tag{20}$$

Equation (19) suggests the interpretation that was given by Jaynes [13] and Grandy [15,16]. The expectation value $\langle \delta f_k(x)\rangle$ of the change $\delta f_k(x)$ is the corresponding generalized work. The remaining part of the change $\delta\langle f_k(x)\rangle$ of the expectation value $\langle f_k(x)\rangle$ comes from the change in the probability distribution $\{p_i\}$ and represents the generalized heat δQ_k for the quantity $f_k(x)$. If the function $f_k(x)$ is such that $f_k(x_i) = E_i$ for all i, then δQ_k is the heat in the usual sense. Grandy [15,16] interpreted equations (19) as the general rule in the probability theory, whose special case is the first law of thermodynamics. Indeed, for a macroscopic system, if $f_k(x_i) = E_i$ for all i, then the corresponding equation (19) has the form of the first law of thermodynamics

$$\delta Q = \delta\langle E\rangle - \langle \delta E\rangle = \delta U - \delta W, \tag{21}$$

where $\langle E\rangle = U$ is the internal energy of the system and $\delta W = \langle \delta E\rangle$ is the work done on the system. According to references [15,16], the heat δQ is the energy transfered through the degrees of freedom over which we don't have control, while the work δW is the energy transfered through the degrees of freedom which we do control. In such an interpretation, the generalized δQ_k is the part of the change of the corresponding expectation value $\delta\langle f_k(x)\rangle$ related to the change in the probability distribution by equation (19). Equations (19) and (20) explictily show that the change in the maximum of information entropy comes from the change in the probability distribution related to δQ_k. Furthermore, Grandy has brought generalized terms δQ_k into connection with the change of the macroscopic constraints brought by means of the external influences on the system. Based on that, Grandy [15,16,19,20] has developed a generalized approach which, along with the generalization of the Liouville-von Neumann equation for the density matrix through the application of the MaxEnt formalism, leads to the derivation of the macroscopic equations of motion.

Let us consider now the quasistatic change of the energy of macroscopic system, for which we specify only that it is a closed system (i.e. the system that can exchange energy, in the form of work or heat, with the environment, but not particles). From equations (20) and (21) then it follows that

$$\delta S = \lambda \delta Q, \tag{22}$$

and

$$\delta U - \delta W = \delta Q = \frac{1}{\lambda}\delta S. \tag{23}$$

If we write the first law of thermodynamics in the form in which the thermodynamic entropy S_e explicitly appears,

$$dU - \delta W = \delta Q = T dS_e, \tag{24}$$

then the Lagrange multiplier λ in the analogous equation (23) can be identified as:

$$\lambda = \frac{1}{kT}. \tag{25}$$

The change δS in the maximum of information entropy given by (22) is thus related to the total differential of thermodynamic entropy dS_e by:

$$k dS = dS_e = \frac{\delta Q}{T}, \tag{26}$$

where T is the temperature, and $1/T$ is the integrating factor for heat δQ. The choice of the unit for temperature (Kelvin), and respectively for entropy (Joule Kelvin^{-1}) is reflected in the appearance of the Boltzmann constant k in the previous expressions.

The confirmation that the identification given by equation (25) is correct comes by introducing the value of the Lagrange multiplier $\lambda = (kT)^{-1}$ in the MaxEnt probability distribution corresponding to the case considered here. In this way we obtain

$$p_i = \frac{1}{Z} \exp\left(-\frac{E_i}{kT}\right), \tag{27}$$

which is known in statistical mechanics as the Gibbs canonical distribution, describing the closed system of known temperature in equilibrium with the environment. The normalization factor of the canonical distribution, the partition function Z, is equal to

$$Z = \sum_{i=1}^{n} \exp\left(-\lambda E_i\right) = \sum_{i=1}^{n} \exp\left(-\frac{E_i}{kT}\right). \tag{28}$$

By considering the open system (i.e. the system that can exchange energy and particles with the environment) in analogous way it is shown that the MaxEnt probability distribution, in the case when along with the expected value of energy, the expected value of the number of particles is known, corresponds to the Gibbs grand canonical distribution [13,15]. Furthermore, it is important that the generic MaxEnt relations from the previous and this section become, in the special cases considered here, the well known equations of equilibrium statistical mechanics.

However, recent work [21] on the Crooks fluctuation theorem [22] and Jarzynski equality [23] indicates further insights. When these important relations of nonequilibrium statistical mechanics are extended to quantum systems strongly coupled with their environments, the thermodynamic entropy of the system of interest in such cases is related to the maximum of the information entropy of the total system (including the system of interest and its environment) minus the information entropy of the environment:

$$S_{e \text{ of the system}} = k\big(S_{I \text{ of the total system}} - S_{I \text{ of the environment}}\big)_{\max}, \tag{29}$$

where k is the Boltzmann constant. The reason for this is that, unlike in the cases considered in this paper, for systems strongly interacting with their environments the correlation between the system and the environment degrees of freedom can not be neglected.

4 MaxEnt and the interpretation of probabilities

In this section we modify and extend the analysis given by Jaynes in reference [17] and show how this leads to the independent interpretation of probabilities which is based on the maximum information entropy principle. Let us consider a proposition $A(n_1, \ldots, n_m)$ which is a function of the sample numbers n_i, $i = 1, 2, \ldots, m$. In the context of statistical mechanics, the sample numbers can represent, for example, the distribution $\{n_1, \ldots, n_m\}$ of the number of systems from the ensemble of $n = \sum_{i=1}^{m} n_i$ identical systems found in m different microscopic states comprising the discrete sample space. The proposition $A(n_1, \ldots, n_m)$ can represent, for example, the expected value of energy of the individual system, or the expected values of some set of compatible quantities. Relative frequencies are then given by $f_i = n_i/n$, $i = 1, 2, \ldots, m$. The number of outcomes for which the proposition A is true is given by the sum over different distributions of sample numbers $\{n_1, \ldots, n_m\}$,

$$M(n, A) = \sum_{\{n_i\} \in R} W(n_1, \ldots, n_m), \tag{30}$$

where R is the region of the sample space for which the proposition A is true and W is the multinomial coefficient

$$W(n_1, \ldots, n_m) = \frac{n!}{n_1! \cdots n_m!}. \tag{31}$$

The greatest term (multiplicity) in the sum (30) over the region R is

$$W_{\max} = \mathrm{Max}_R W(n_1, \ldots, n_m). \tag{32}$$

If $T(n, m)$ is the number of terms in the sum (30), then it is true that

$$W_{\max} \le M(n, A) \le W_{\max} T(n, m), \tag{33}$$

and

$$\frac{1}{n} \log W_{\max} \le \frac{1}{n} \log M(n, A) \le \frac{1}{n} \log W_{\max} + \frac{1}{n} \log T(n, m). \tag{34}$$

From combinatorial arguments it follows that

$$T(n, m) = \binom{n + m - 1}{n} = \frac{(n + m - 1)!}{n!(m - 1)!}. \tag{35}$$

Then as $n \to \infty$

$$T(n, m) \sim \frac{n^{m-1}}{(m - 1)!}. \tag{36}$$

Therefore, as $n \to \infty$, $\log T(n,m)$ grows less rapidly than n,

$$\frac{1}{n}\log T(n,m) \to 0, \tag{37}$$

and from equations (34) and (37) it follows that

$$\frac{1}{n}\log M(n,A) \to \frac{1}{n}\log W_{\max}, \tag{38}$$

as $n \to \infty$. The multinomial coefficient W grows so rapidly with n that the maximum term $W_{\max}$ dominates, in the sense given by equation (38), the total multiplicity $M(n,A)$ given by the sum (30).

However, the limit we really want is the one in which the sample frequencies n_i/n tend to certain (but not yet specified) constant values f_i as $n \to \infty$. Therefore, we want the limit of

$$\frac{1}{n}\log W = \frac{1}{n}\log\left[\frac{n!}{(nf_1)!\cdots(nf_m)!}\right], \tag{39}$$

as $n \to \infty$. Using the Stirling asymptotic approximation

$$\log n! \sim n\log n - n + \log\sqrt{2\pi n} + O\left(\frac{1}{n}\right), \tag{40}$$

we find as $n \to \infty$ that in this limit we have

$$\frac{1}{n}\log W \to H \equiv -\sum_{i=1}^{m} f_i \log f_i, \tag{41}$$

and this gives the information entropy of the relative frequency distribution $\{f_1,\ldots,f_m\}$. So, from equations (38) and (41), it follows that in a such limit we also have that

$$\frac{1}{n}\log M(n,A) \to \frac{1}{n}\log W_{\max} = H_{\max}. \tag{42}$$

Therefore, for very large n, the maximum multiplicity $W_{\max}$ is the one that dominates the total multiplicity $M(n,A)$ and maximizes the information entropy H subject to the constraints that define the region of the sample space for which the proposition A is true. Furthermore, it is straightforward to show that the probability of obtaining the relative frequency distribution $\{f_1,\ldots,f_m\}$ which corresponds to the maximum multiplicity $W_{\max}$ approaches 1 in the limit of large n, because from equation (38) in this limit we have

$$\frac{W_{\max}}{M(n,A)} \sim 1. \tag{43}$$

Therefore, in the limit of large n, without any other additional constraints except that the proposition A is true, we can assume with certainty that the relative frequencies to be used are the ones that maximize the multiplicity W and, because of equations (41) and (42), maximize the information entropy H. Therefore, according to the weak law of large numbers, the relative frequencies in the limit of a large number of trials ($n \to \infty$) should correspond to the MaxEnt probabilities.

So, in this context, can we now examine the frequency interpretation of probabilities as factual properties of the real world, if, as in this example, the corresponding probabilites (obtained in the limit of a large number of trials) actually follow from the principle of maximum information entropy, and therefore, depending only on the available information (i.e. on the proposition A), depend on our state of knowledge? This question comes naturally as the above result implies that under constraints representing the information that is available, the relative frequencies in the limit of a large number of trials tend with certainty to the corresponding MaxEnt probabilities.

5 Conclusion

We have shown how the probabilities in statistical mechanics can not be simply interpreted in the frequentist context. Probabilities, at least in the Gibbs formalism of statistical mechanics, are not simply relative frequencies in the ensemble of a large number of identical systems. Actually they depend on the available information about the individual system and therefore are the description of a degree our knowledge about it. The ensembles of identically prepared systems are chosen in the Gibbs formalism only to illustrate that the information we have about the individual system is incomplete, which means that it is not sufficiently detailed to specify the exact microscopic state of a macroscopic system, nor its exact evolution in time.

Furthermore, in the case of nonequilibrium systems and processes that are irreversible on the macroscopic level, justification of nonequilibrium ensembles in the frequentist sense as a physical fact, using only first principles, via equations of motion and ergodic theorems, becomes permeated with technical and, more importantly, conceptual difficulties [24]. For example, the applications of ergodic theorems for that purpose would require an infinite or large time intervals, and this is not in general always available for nonequilibrium systems that are continuously evolving and changing its macroscopic state with time. This is well exemplified in the work of Zubarev and his coworkers, who introduce a hierarchy of time scales, with different sets of quantities that are relevant for the description of a nonequilibrium system on different time scales [25,26]. More important than ergodicity is the concept of a mixing system, originally introduced by Gibbs [8]. Mixing implies ergodicity, and hopefully can provide a mechanical foundation of both nonequilibrium and equilibrium statistical mechanics, if we can prove it for realistic systems [6]. However, there are differing opinions about its importance since transport coefficients and dissipativity, an essential property of macroscopic systems, can not be derived only from mixing [27]. On the other hand, as we have shown here, MaxEnt formalism is an independent logical extension of the Gibbs method, and leads to statistical distributions which depend only on the available information. If that information is relevant for the description of a system at a macroscopic level, then accordingly, the obtained statistical distributions should be relevant for describing its macroscopic state, its properties and time evolution [7,10,13–16,18–20,25,26,28–30].

References

1. E.T. Jaynes, Predictive statistical mechanics, in *Frontiers of Nonequilibrium Statistical Physics*, edited by G.T. Moore, M.O. Scully (Plenum Press, New York, 1986), pp. 33−56.
2. W. Feller, *An Introduction to Probability Theory and Its Applications* (John Wiley and Sons, New York, 1961)
3. Law of large numbers, in *Encyclopedia of Mathematics*. http://www.encyclopediaofmath.org/
4. O. Penrose, *Foundations of Statistical Mechanics* (Pergamon Press, Oxford, 1970)
5. I. Farquhar, *Ergodic Theory in Statistical Mechanics* (Wiley, New York, 1964)
6. J.R. Dorfman, *An Introduction to Chaos in Nonequilibrium Statistical Mechanics* (Cambridge University Press, Cambridge, 1999)
7. E.T. Jaynes, Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, edited by R.D. Levine, M. Tribus (MIT Press, Cambridge, 1979), pp. 15−118
8. J.W. Gibbs, *Elementary Principles in Statistical Mechanics* (Yale University Press, New Haven, 1902)
9. E.T. Jaynes, Am. J. Phys. **33**, 391 (1965)
10. E.T. Jaynes, Macroscopic prediction, in *Complex Systems Operational Approaches in Neurobiology, Physics, and Computers*, edited by H. Haken (Springer, Berlin, 1985), pp. 254−269.
11. E.T. Jaynes, The second law as physical fact and as human inference, Unpublished manuscript (1990), http://bayes.wustl.edu/etj/node2.html
12. C.E. Shannon, Bell Syst. Tech. J. **27**, 379, 623 (1948). Reprinted in *The Mathematical Theory of Communication*, edited by C.E. Shannon, W. Weaver (University of Illinois Press, Urbana, 1949)
13. E.T. Jaynes, Phys. Rev. **106**, 620 (1957)
14. E.T. Jaynes, Phys. Rev. **108**, 171 (1957)
15. W.T. Grandy, *Entropy and the Time Evolution of Macroscopic Systems* (Oxford University Press, Oxford, 2008)
16. W.T. Grandy, Found. Phys. **34**, 1 (2004)
17. E.T. Jaynes, in *Probability Theory: The Logic of Science*, edited by G.L. Bretthorst (Cambridge University Press, Cambridge, 2003)
18. D. Kuić, Foundational principles of predictive statistical mechanics as the basis for theory of irreversibility, Ph.D. thesis (in Croatian) (2013), http://digre.pmf.unizg.hr/22/
19. W.T. Grandy, Found. Phys. **34**, 21 (2004)
20. W.T. Grandy, Found. Phys. **34**, 771 (2004)
21. M. Campisi, P. Talkner, P. Hanggi, Phys. Rev. Lett. **102**, 210401 (2009)
22. G.E. Crooks, Phys. Rev. E **60**, 2721 (1999)
23. C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997)
24. E.T. Jaynes, Foundations of probability theory and statistical mechanics, in *Delaware Seminar in the Foundations of Physics*, edited by M. Bunge (Springer, Berlin, 1967), pp. 77−101
25. D. Zubarev, V. Morozov, G. Ropke, *Statistical Mechanics of Nonequilibrium Processes, Vol. 1: Basic Concepts, Kinetic Theory* (Akademie Verlag, Berlin, 1996)
26. D. Zubarev, V. Morozov, G. Ropke, in *Statistical Mechanics of Nonequilibrium Processes, Vol. 2: Relaxation and Hydrodynamic Processes* (Akademie Verlag, Berlin, 1997)
27. R. Balescu, *Equilibrium and Nonequilibrium Statistical Mechanics* (John Wiley and Sons, New York, 1975)
28. D. Kuić, P. Zupanovic, D. Juretic, Found. Phys. **42**, 319 (2012)
29. D. Kuić, Predictive statistical mechanics and macroscopic time evolution. A model for closed Hamiltonian systems, arXiv:1506.02622 (2015)
30. D. Kuić, Predictive statistical mechanics and macroscopic time evolution. Hydrodynamics and entropy production. Accepted for publication in Found. Phys, arXiv:1506.02625 (2015)