# Combinatorial Information Theory:
# I. Philosophical Basis of Cross-Entropy and Entropy

Robert K. Niven[1]

[1]*School of Aerospace, Civil and Mechanical Engineering,*
*The University of New South Wales at ADFA,*
*Northcott Drive, Canberra, ACT, 2600, Australia.*[*]

(Dated: 9 January 2006)

## Abstract

The three main theoretical bases of the concepts of entropy and cross-entropy - information-theoretic, axiomatic and combinatorial - are critically examined. It is shown that the combinatorial basis, proposed by Boltzmann and Planck, is the most fundamental (most primitive) basis of these concepts, since it provides (i) a derivation of the Kullback-Leibler cross-entropy and Shannon entropy functions, as simplified forms of the multinomial distribution subject to the Stirling approximation; (ii) an explanation for the need to maximize entropy (or minimize cross-entropy) to find the most probable realization; and (iii) the means to derive entropy and cross-entropy functions for systems which do not satisfy the multinomial distribution, i.e. which fall outside the domain of the Kullback-Leibler and Shannon measures. The information-theoretic and axiomatic bases of cross-entropy and entropy - whilst of tremendous importance and utility - are therefore seen as secondary viewpoints, which lack the breadth of the combinatorial approach. Appreciation of this reasoning would permit development of a powerful body of "combinatorial information theory", as a tool for statistical inference in all fields (inside and outside science). The essential features of Jaynes' analysis of entropy and cross-entropy - reinterpreted in light of the combinatorial approach - are outlined, including derivation of probability distributions, ensemble theory, Jaynes relations, fluctuation theory and Jaynes' entropy concentration theorem. New results include a generalized free energy (or "free information") concept, a generalized Gibbs-Duhem relation and phase rule. Generalized (combinatorial) definitions of entropy and cross-entropy, valid for any combinatorial system, are then proposed and examined in detail.



PACS numbers: 02.50.Cw, 02.50.Tt, 05.20.-y, 05.20.Gg, 05.30.-d, 05.30.Ch, 05.40.-a, 05.70.-a, 05.70.Ce, 05.90.+m, 64.10.+h, 89.20.-a, 89.70.+c,

Keywords: entropy, cross-entropy, directed divergence, probability, information theory, bits, axiomatic, combinatorial, Boltzmann principle, thermodynamics, statistical mechanics, free energy, Jaynes, maximum entropy, minimum cross-entropy, statistical inference



*Electronic address: r.niven@adfa.edu.au; Telephone: +61-2-6268-8330; Fax: +61-2-6268-8337

## I. INTRODUCTION

Since its inception one and a half centuries ago, the concept of *entropy* has been mired in controversy, and remains the subject of widespread confusion. After a period of gestation in the 1850s by Clausius, Kelvin, Rankine and many others, the *thermodynamic entropy* was formally defined by Clausius [1] in terms of the exact differential $dS$, given by the quantity of heat transferred reversibly to a system $dQ$ scaled by the absolute temperature $T$ of the system:

$$dS = \frac{dQ}{T} \tag{1}$$

Consideration of irreversible (non-equilibrium) processes - as expressed by the second law of thermodynamics - gives the Clausius [2] inequality:

$$dS \geq \frac{dQ}{T} \tag{2}$$

Boltzmann [3] and Planck [4, 5] expounded the statistical basis of entropy, based on the quantization of matter, giving for molecular systems:

$$S_N = NS = k \ln \mathbb{W} \tag{3}$$

where $S_N$ is the total thermodynamic entropy of the system, $N$ is the number of entities (discrete particles or agents) present, $S$ is the thermodynamic entropy per entity, $\mathbb{W}$ is the statistical weight or number of possible realizations[1] of the system, of equal probability, and $k$ is the Boltzmann constant ($1.38 \times 10^{23}$ J K$^{-1}$ entity$^{-1}$). For discrete systems, (3) has been given as:

$$S = -k \sum_{i=1}^{s} p_i \ln p_i \tag{4}$$

where $p_i$ is the probability of occurrence of the $i$th distinguishable outcome or state (e.g. the $i$th energy level), from a total of $s$ such states. The thermodynamic entropy is the object of the second and third laws of thermodynamics, the Clausius inequality and the Helmholtz and Gibbs free energy functions, with widespread application to the analysis of physical, chemical and energetic systems.

[1] Here the *state* refers to each different category (e.g. boxes, energy levels, elements or results) accessible to a system, whilst the *realization* is the actual physical pattern of the system amongst its states (*complexion*, *microstate* or *outcome*).

Shannon [6] initiated the field of information theory by the introduction of the *information entropy*[2]:

$$H(\mathbf{p}) = -\sum_{i=1}^{s} p_i \ln p_i \tag{5}$$

where $\mathbf{p} = \{p_i\}$ is the set of all $p_i$. (Often $H$ is multiplied by a scaling constant, $K$, here taken as unity.) Semantically, the Shannon entropy differs from the thermodynamic entropy in that it is more generally based on information theory (as will be discussed), and is broadly applicable to many different types of systems. The Shannon entropy therefore encompasses the thermodynamic entropy concept as a special case [7, 8, 9, 10, 11, 12]. As it includes the scaling factor $k$ and is therefore dimensional, the $S$-entropy may also be referred to as the *scaled entropy*. The *maximum entropy position* of a system is considered to have the most uncertainty, is least biased, preserves the least information, or is least committed to the information not given [8, 10, 12]. Thus for a system described by $x$ equations in $y$ unknowns with $y > x$, the *maximum entropy principle* ("MaxEnt") provides a formal mechanism for predicting the expected probability distribution of the governing variable, subject *only* to what is known. Jaynes [8] and Tribus [9, 10] demonstrated that all thermodynamics can be derived directly (and more naturally) from the maximum entropy principle without recourse to any other laws. The MaxEnt concept has found widespread application to virtually all fields of human endeavour, including information technology, communications, mathematics, science, engineering, economics, decision theory, geography, linguistics and the social sciences (e.g. [10, 11, 12, 13, 14]).

The Shannon information entropy is itself a subset of the Kullback-Leibler *directed divergence* or *cross-entropy* function [12, 15, 16, 17], which is in discrete form:

$$D(\mathbf{p}|\mathbf{q}) = \sum_{i=1}^{s} p_i \ln \frac{p_i}{q_i} \tag{6}$$

where $q_i$ and $p_i$ are respectively the *a priori* and *a posteriori* probability of occurrence of the $i$th result, $\mathbf{p} = \{p_i\}$, $\mathbf{q} = \{q_i\}$ and the solidus | is used in the Bayesian sense to indicate "subject to". Note $-D$ is also referred to as the *relative entropy* [18]. The cross entropy provides a means of measuring the probabilistic "distance" of the probability distribution $\mathbf{p}$ from $\mathbf{q}$. Minimization of $D$ subject to the constraints upon a system ("MinXEnt") yields the

[2] The standard symbol $H$ is unfortunate, since it clashes with that for enthalpy and the Hamiltonian; to avoid confusion, these terms are not referred to here.

distribution $\mathbf{p}$ which satisfies these constraints, yet is closest to $\mathbf{q}$ [12]. When $\mathbf{q}$ is the uniform distribution, $\mathbf{u}$ (ie, $q_i = 1/s = u =$ constant for all $i$), minimization of the cross-entropy yields the negative of the Shannon entropy, shifted by a constant [12].

A variety of other entropy, information and divergence functions abound, for example the continuous Shannon entropy [6]; von Neumann entropy [19]; Fisher information [20, 21]; Rényi entropy [22]; Kolmogorov entropy [23]; Tsallis entropy [24, 25]; Kaniadakis entropy [26, 27] and many others (e.g. [14, 28, 29, 30, 31, 32, 33, 34]). Such variants are not as prominent as the Shannon entropy, but have application to information theory in its most general sense or to specific applications. They are not discussed further here except where relevant [3].

The aims of this and the following study (Parts I and II), which extend previous studies [35, 36], are twofold. In Part I, the three main theoretical roots of the Shannon information entropy and Kullback-Leibler cross-entropy concepts - information-theoretic, axiomatic and combinatorial - are examined, leading into an analysis of the traditional maximum entropy (MaxEnt) derivation of the generalized Maxwell-Boltzmann distribution, and an equivalent derivation based directly on combinatorial principles. It is shown (following a well-trodden road) that both the cross-entropy and entropy functions are simplified forms of the logarithm of the multinomial distribution; they are therefore only shorthand functions to determine the most probable (minimum cross-entropy or maximum entropy) realization of a system which follows the multinomial distribution, without the necessity of invoking this distribution itself. The Kullback-Leibler cross-entropy and Shannon information entropy functions are therefore secondary concepts, based firmly on simple combinatorial principles. This perspective lies in stark contrast to the axiomatic philosophical basis now dominant in the information literature (e.g. [6, 8, 12, 37]), which sees the cross-entropy or (especially) the entropy function as the fundamental basis and starting point for analysis. It also opens the door to different cross-entropy and entropy functions, applicable to systems (e.g. Bose-Einstein, Fermi-Dirac, Rényi, Tsallis, Kaniadakis, etc) which do not follow the multinomial distribution. Note that much of this analysis is not new, but encompasses and expands upon the philosophical arguments of statistical mechanics (e.g. [3, 4, 5, 7, 38, 39, 40, 41, 42, 43, 44, 45, 46]), which are examined only in passing by information theorists (notable exceptions include

[3] The reader will appreciate the irony in the proliferation of many different entropy functions.

[17, 18, 47, 48, 49, 50]). Appreciation of this reasoning would permit development of a much more powerful body of "combinatorial information theory", applicable to problems outside the scope of the Shannon and Kullback-Leibler measures.

For completeness, the main features of Jaynes' [8, 18, 47, 48] and later workers' treatments (e.g. [9, 10, 12, 14, 50]) of the maximum entropy and minimum cross-entropy methods are reproduced and extended using "first combinatorial principles". This includes the derivation and discussion of probability distributions, ensemble theory, Jaynes relations, a generalized free energy (or "free information") function, Gibbs-Duhem relation, phase rule, fluctuation theory, and Jaynes' entropy concentration theorem. Generalized definitions of entropy and cross-entropy, applicable to any combinatorial system, are also provided.

Part II examines an important oversight in the usual definitions of cross-entropy and entropy, as applied to quantized systems: the assumption that the total number of entities or trials, $N$, and/or the numbers of entities or selections in each category, $n_i$, approach infinity. This assumption is inherent in the Stirling [51]-de Moivre [52] approximation, applied almost universally throughout statistical mechanics. Whilst appropriate for *well-populated systems*, such as physical or chemical systems containing of the order of Avogadro's number ($6.02 \times 10^{23}$) of entities, in the case of *sparsely populated systems* (including quantum mechanical systems) it is not. Using combinatorial principles, the exact forms of the Kullback-Leibler cross-entropy and Shannon entropy functions, referred to as the *exact cross-entropy* and *exact entropy* respectively, are derived from the multinomial distribution without use of Stirling's approximation. Examination of these functions reveals some surprising properties, including non-additivity. The exact forms of the Maxwell-Boltzmann distribution and cross-distribution, and of a number of "Jaynes relations" [8, 18, 48], are also obtained. A theory of *exact thermodynamics* is then built up in the manner of Jaynes [8], revealing that sparsely populated multinomial systems satisfy (on average) the zeroth, first, second and third laws of thermodynamics. However, fluctuations from the maximum entropy position become much more important. The analysis provides a more detailed theoretical framework for recent studies of the exact Maxwell-Boltzmann and other entropy functions [35, 36].

In the following analysis, an *entity* is taken to be any discrete particle, object or agent within a system, which acts separately but not necessarily independently of the other entities present (note this definition encompasses human beings). The entity therefore constitutes the unit of analysis of the system, although of course some entities can be further examined

in terms of their constituent sub-entities, if desired. A *well-populated system* is one in which the total number of entities, $N$, and the number of entities in each category, $n_i$, approach infinity, as distinct from a *sparsely populated system*, in which they do not. Parts I and II primarily concern *multinomial* systems, i.e. those governed by the multinomial distribution, which (as will be shown) encompasses Maxwell-Boltzmann statistics.

## II. THEORETICAL ROOTS OF THE INFORMATION ENTROPY CONCEPT

What is entropy? This question has certainly occupied (or been dismissed from) the minds of millions of college and university students for one and a half centuries - predominantly in physics, chemistry, engineering and informatics - and undoubtedly tens of thousands more of their professional elders in all disciplines. To endeavour to answer this question, in this section the three primary theoretical or philosophical roots of the entropy concept - or more specifically, of the information entropy and cross-entropy functions - are examined. The first two, information-theoretic and axiomatic, are so closely intertwined in the literature that it is not possible to distinguish them clearly. The third origin, based on combinatorial analysis, is somewhat distinct. Discussion of a fourth origin, involving the inverse methods of Kapur, Kesavan and co-workers [11, 12, 53, 54, 55, 56], is postponed until later in the text (section III B). A fifth origin based on game theory, as proposed by Topsøe [57, 58], is also discussed. A rival approach to the analysis of probabilistic systems, which invokes the continuous Fisher information [20, 21, 59] is not discussed here, but will be scrutinized elsewhere.

### A. The Information-Theoretic (Bits) Approach

The first theoretical basis of the Shannon entropy - although not the first in historical development - concerns the number of bits of information required to specify a particular system or outcome [6, 37, 60, 61, 62, 63]. Consider the *binary entropy* or $B$-entropy:

$$B = -\sum_{i=1}^{s} p_i \log_2 p_i \tag{7}$$

related to the Shannon entropy (defined using the natural logarithm, (5)) by $H = B \ln 2$. Now consider a random variable which may take one of two states, of equal probability

$p_i = \frac{1}{2}, i = 1, 2$. Initially, the state of the variable is not known. After a *binary decision* (a process of selection or measurement) it is found to be in one of these states (say $p_1 = 1$) and not the other ($p_2 = 0$). The initial and final binary entropies are therefore:

$$B_{init} = -2(\frac{1}{2}\log_2\frac{1}{2}) = 1, \qquad B_{final} = -(1\log_2 1 + 0\log_2 0) = 0 \tag{8}$$

(Here and subsequently, we take $0\log 0 = \log 0^0 = \log 1 = 0$ for all logarithmic bases). The change in entropy is then:

$$\Delta B = B_{final} - B_{init} = -1 \tag{9}$$

If we *define* the change in information as the negative of the change in entropy (i.e., entropy lost = information gained) [39, 61, 62, 63, 64, 65], the gain in information - reflecting our improved state of knowledge - is:

$$\Delta I = -\Delta B = 1 \tag{10}$$

Thus for a simple binary decision, the information gained (entropy lost) corresponds to one *bit* of information. The decrease in entropy therefore provides a quantitative measure of the information gained by observation of a system.

If we adopt a scaled binary entropy $S_B = -k\sum_{i=1}^{s} p_i \log_2 p_i$, the information gained by a binary decision is $k$, measured in the units of $k$. For a scaled entropy based on the natural logarithm, $S = -k\sum_{i=1}^{s} p_i \ln p_i$, the gain in information is $k\ln 2$ [6, 60]. For thermodynamic systems for which $k$ is the Boltzmann constant, 1 bit of information corresponds to an energy transfer of $9.57 \times 10^{-24}$ J K$^{-1}$ entity$^{-1}$. To access information carried by photons, and distinguish them from the background (thermal) radiation, it is necessary to account for the effect of temperature [62, 63]; in this case, 1 bit of information corresponds to $kT\ln 2$ energy units per entity.

A second variant of the information-theoretic definition - which overlaps with the axiomatic approach (section II B) - is to consider a random variable which may take $s$ equally probable states. We define a measure of *uncertainty* as [10, 66]:

$$U = \ln s \tag{11}$$

As the states are equally probable, $s = 1/p_i, \forall i$, hence $U = -\ln p_i$. The mathematical expectation of the uncertainty is $\langle U\rangle = -\sum_{i=1}^{s} p_i \ln p_i = H$, i.e. the Shannon entropy. As the states are equally probable, this reduces to $\langle U\rangle = U$.

For states which are not equally probable, we may thus adopt the Shannon entropy as a measure of the expectation of the uncertainty [6]. We can further define the *surprisal* or *self-information* associated with each result [9, 10, 29]:

$$\sigma = -\ln p_i \tag{12}$$

The entropy is therefore the expectation of the surprisal.

The surprisal has also been defined relative to the *a priori* probability of that result, $\delta = \ln(p_i/q_i)$, i.e. as the amount of information gained by a decision or message [10, 15, 29]. This is better referred to as the *cross-surprisal*. The expectation of the cross-surprisal gives the cross-entropy (6). The cross-entropy is therefore a measure of the expected information relative to what is known.

Another useful term is the function $H_i = -p_i \ln p_i$, here termed the *weighted surprisal* or *partial entropy*, which when summed over all states gives the Shannon entropy (c.f. [67, 68, 69, 70]). The analogous function $D_i = p_i \ln(p_i/q_i)$ can be termed the *weighted cross-surprisal* or *partial cross-entropy*.

Whilst of great utility, the above information-theoretic roots of the Shannon entropy both suffer from the deficiency that they *assume* that measures of information (or entropy) should be of logarithmic form, an assumption in part derived from the axiomatic approach (section II B). Certainly, other functions could yield one bit of information for a binary decision (10), whilst it is not at all clear why either the uncertainty or surprisal should be logarithmic (11-12). Some authors have tried to justify this choice on the grounds of human perception of physical stimuli (see [34]), a rather unsatisfactory explanation. To address this question it is first necessary to consider the axiomatic approach, now the dominant theoretical or philosophical root of the Shannon entropy and Kullback-Leibler cross-entropy functions.

### B. The Axiomatic Approach

The second theoretical basis of the entropy concept, developed by Shannon (1948), proceeds by listing the desired properties of a measure of uncertainty - its *axioms* or *desiderata* - and finding the mathematical function which satisfies these axioms. Shannon (1948) considered three axioms: continuity, monotonicity and recursivity (the branching principle),

from which the Shannon entropy (5) is uniquely obtained. To Shannon's original list, many additional axioms have been added: e.g. uniqueness, permutational symmetry (invariance), non-negativity, non-impossibility, inclusivity, decisivity, concavity, maximum entropy at uniformity (normality), additivity, strong additivity, subadditivity, system independence and subset independence (e.g. [6, 10, 12, 14, 28, 31, 33, 71]). The Shannon entropy is the only function which satisfies these axioms. Indeed, it may be deduced from several small subsets of these axioms, implying that they are not independent (e.g. [12, 31, 72]).

It must be noted that Planck's [5] definition of thermodynamic entropy (3) is derived by an axiomatic argument, assuming multiplicity of the weights and additivity of the entropy function. Similarly, in the "plausible reasoning" treatises of Cox [73: p37] and Jaynes [48: section 2.1], the Shannon entropy (5) is obtained axiomatically, assuming entropy is additive and multiply differentiable.

The cross-entropy or directed divergence function $D$ can also be obtained using the axiomatic approach [12, 15, 16, 71]. Its governing axioms are broadly similar to those for the Shannon entropy, except that it is convex, and the equilibrium distribution $\mathbf{p}^* = \mathbf{q}$ in the absence of other constraints [12]. Both the maximum entropy and minimum cross-entropy principles have also been justified axiomatically (e.g. [71, 74]).

Whilst mathematically sound and of tremendous utility, the axiomatic approach is intellectually unsatisfying in that it presents an austere, sterile basis for the entropy and cross-entropy functions, based only on abstract notions of desirable properties. The answer to the question - what is entropy? - is still not clear. Further, as Kapur [31: p209] notes: "*mathematicians tried to modify these axioms to get more general measures* [of uncertainty] *including Shannon's measure as a special or limiting case*". Other entropy functions, which do not reduce to the Shannon entropy, have also been derived using different sets of axioms (e.g. [14, 22, 24, 25, 28, 31, 33, 34]). Other measures of divergence have also been proposed (e.g. [29, 30, 32, 34]). How can we be certain that the axioms used to derive the Shannon or Kullback-Leibler measures are correct? Indeed, the specification of particular axioms may preclude the identification of different or broader measures of entropy, which may be more appropriate for particular or more general circumstances. To resolve these circular arguments, we now turn to consideration of the combinatorial basis of the entropy and cross-entropy functions, which as will be shown, should be recognized as their primary (most primitive) philosophical basis.

### C. The Combinatorial (Statistical Mechanical) Approach

#### 1. *Statistics of Multinomial Systems*

The combinatorial approach was first developed in statistical thermodynamics, to examine the distribution of molecules amongst energy levels or phase space elements (e.g. [3, 4, 5, 7, 38, 39, 40, 41, 42, 43, 44, 45, 46]). However, the combinatorial basis is only touched upon by many prominent statistical mechanics texts (e.g., [75]) in favour of a quantum mechanical treatment, which tends to disguise its statistical foundation. The connection between combinatorial concepts and entropy is not prominent in the information theory literature, although there are a number of notable exceptions (e.g. [17, 18, 47, 48, 49, 50]).

Consider the "balls-in-boxes" system illustrated in Figure 1a, in which $N$ distinguishable balls or entities are distributed amongst $s$ distinguishable boxes or states. This may be taken to represent $N$ molecules amongst $s$ energy levels, phase space elements or eigenfunctions[4] ; $N$ ensemble members amongst $s$ ensemble energy values; $N$ people amongst $s$ shops; $N$ cars amongst $s$ floors of a parking station, and so on. Each realization of the system will contain $n_1$ balls in box 1, $n_2$ balls in box 2, etc, or in general $n_i$ balls in box $i$. The $N$ balls are taken to be distinguishable, but their permutations within each box are indistinguishable, i.e. we can only (or need only) distinguish the balls within any given box from those in the other boxes. Each choice (of a ball in a box) is assumed independent of the other selections. The probability of any particular realization of the system, $\mathbb{P}$ (equal to the probability that there are $n_i$ balls in the $i$th box, for each $i$), is given by the *multinomial distribution* [76, 77, 78]:

$$\mathbb{P}|\mathbf{q} = \frac{N!}{n_1!n_2!...n_s!} q_1^{n_1} q_2^{n_2} ... q_s^{n_s} = N! \prod_{i=1}^{s} \frac{q_i^{n_i}}{n_i!} \tag{13}$$

where again $q_i$ is the *a priori* probability of a ball falling in the $i$th box. If $\mathbf{q} = \mathbf{u}$ (i.e. $q_i = u = 1/s, \forall i$) this reduces to:

$$\mathbb{P}|\mathbf{u} = \frac{N!}{\prod_{i=1}^{s} n_i!} s^{-N} \tag{14}$$

[4] The boxes are here taken to be discrete, although there is no conceptual difficulty in generalizing the analysis to boxes of infinitesimal spacing. Similarly, the number of states $s$ is considered finite, but the limit $s \to \infty$ can be considered if handled carefully [48].

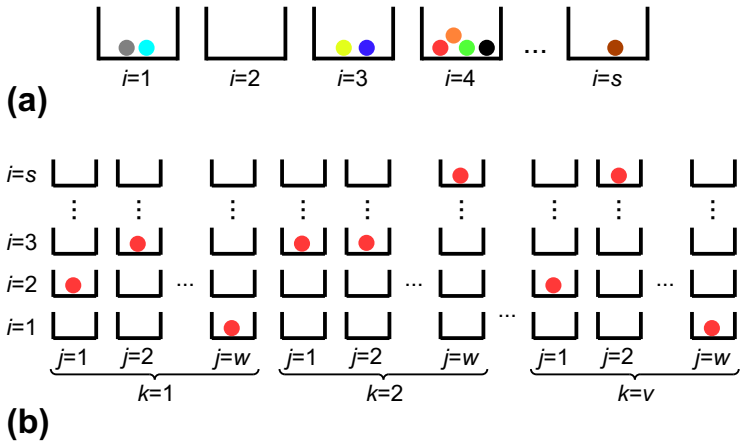


FIG. 1: Multinomial (a) balls-in-boxes and (b) multiple selection systems ("color online").

Since the total number of permutations of a multinomial distribution is $s^N$ [79], the number of ways in which any particular realization in (14) can be produced, or its statistical *weight*, is [80, 81]:

$$\mathbb{W} = (\mathbb{P}|\mathbf{u})\, s^N = \frac{N!}{\prod_{i=1}^{s} n_i!} \tag{15}$$

For constant $N$, the above equations are subject to the *natural constraint*:

$$\text{C0}: \qquad \sum_{i=1}^{s} n_i = N \tag{16}$$

and usually one or several *moment constraints* (c.f. [8]):

$$\text{C1 to C}R: \qquad \sum_{i=1}^{s} n_i f_{ri} = N \langle f_r \rangle, \qquad r = 1, ..., R \tag{17}$$

where $f_{ri}$ is the value of the function $f_r$ in the $i$th state and $\langle f_r \rangle$ is the mathematical expectation of $f_{ri}$. An example of (17) is an energy constraint, in which each state is of energy $f_{1i} = \varepsilon_i$ and the expectation of the energy is $\langle f_1 \rangle = \langle \varepsilon \rangle$.

Now consider a sequence of $v$ independent, identical probabilistic *events*, within each of which $w$ trials or *selections* are made between $s$ possible states or *results*, as represented in Figure 1b. Examples include tosses of a coin or coins, throws of a die or dice, spins of a roulette wheel, choices of symbols to make up a communications signal, or the sexual liaisons of leading film star. So long as we are only interested in the statistical nature of the selections, and not their order, the probability of any realization (without regard to order, assuming each event is independent) also follows the multinomial distribution (13)

with $N = vw$. When only one selection is made in each event (i.e. $w = 1$), then $N = v$. When the *a priori* probabilities $q_i$ of each state within each selection are identical, the weight also follows (15).

The above "balls-in-boxes" and "multiple selection" systems (Figures 1a-b) are both examples of *multinomial systems*, i.e. those in which the probability of any given realization, $\mathbb{P}$, follows the multinomial distribution.

### *2. The Most Probable Realization*

We now use *first combinatorial principles* to determine the *most probable realization* of the multinomial systems considered. As mentioned, the following derivation is common in statistical thermodynamics (e.g. [3, 4, 5, 7, 38, 39, 40, 41, 42, 43, 44, 45, 46]), although such workers base their derivations on the weight $\mathbb{W}$. As it is based on $\mathbb{P}$ rather than $\mathbb{W}$, the following derivation incorporates the *a priori* probabilities $\mathbf{q}$, and is therefore more comprehensive [50].

Clearly, the most probable realization is that for which $\mathbb{P}$ (13) is a maximum, subject to the constraints C0-C$R$ on the system ((16), (17)). As the natural logarithm $\ln x$ increases monotonically with $x$, but transforms a product into a sum, it is convenient - and equivalent - to maximize $\ln \mathbb{P}$ rather than $\mathbb{P}$ itself, a convention adopted (implicitly) throughout statistical mechanics [3, 46]. (The use of logarithms is therefore merely a matter of convenience, not a requirement.) The most probable realization is given by:

$$d\left(\ln \mathbb{P} | constraints\right) = 0 \tag{18}$$

where $d(\ )$ is the total derivative or variational operator. Note (18) can be constructed using Lagrange's method of undetermined multipliers [10, 12, 46], involving extremization of the Lagrangian $\mathfrak{L}$:

$$d\mathfrak{L} = 0 \tag{19}$$

From the multinomial distribution (13):

$$\ln \mathbb{P} = \sum_{i=1}^{s} \left( \frac{n_i}{N} \ln N! - \ln n_i! + n_i \ln q_i \right) \tag{20}$$

in which (for reasons which will become clear in Part II) the leading $\ln N!$ term is brought inside the summation using the natural constraint (16). From the constraints (16)-(17), the

Lagrangian is:

$$\mathfrak{L} = \sum_{i=1}^{s} \left( \frac{n_i}{N} \ln N! - \ln n_i! + n_i \ln q_i \right) - (\lambda_0 - 1) \left( \sum_{i=1}^{s} n_i - N \right) - \sum_{r=1}^{R} \lambda_r \left( \sum_{i=1}^{s} n_i f_{ri} - \langle f_r \rangle N \right) \tag{21}$$

where $\lambda_r$, $r = 0, ..., R$, are the Lagrangian multipliers, and $\lambda_0 - 1$ is chosen rather than $\lambda_0$ for mathematical convenience. For constant $N$, $q_i$ and $\langle f_r \rangle$, and for $f_{ri}$ independent of $n_i$, we need only consider the variation of (21) with respect to $n_i$, i.e. $\partial \mathfrak{L} / \partial n_i \; dn_i = 0, \forall i$, whence:

$$\frac{1}{N} \ln N! - \frac{\partial}{\partial n_i} \ln n_i! + \ln q_i - (\lambda_0 - 1) - \sum_{r=1}^{R} \lambda_r f_{ri} = 0, \qquad i = 1, ..., s \tag{22}$$

The above equations are expressed in terms of $n_i$, and can be said to be in "$n_i$ form."

At this stage the near-universal approach taken in the literature (see previous statistical mechanics references) is to employ a truncated form of the approximation for factorials derived by Stirling [51] and de Moivre [52]:

$$\ln x! \approx x \ln x - x \tag{23}$$

This is accurate to within 1% of $\ln x!$ for $x > 90$. (A more precise form, $\ln x! \approx x \ln x - x + \frac{1}{2} \ln(2\pi x)$, is accurate to within 1% of $\ln x!$ for $x > 4$ [77]). Thus $\partial \ln n_i! / \partial n_i \approx \ln n_i$ and $\ln N! \approx N \ln N - N$, and so the most probable realization, here designated with an asterisk, is obtained from (22) in conjunction with C0 (16) as (c.f. [8, 9, 10, 12, 14]):

$$n_i^* | q_i \approx N q_i \exp \left( -\lambda_0 - \sum_{r=1}^{R} \lambda_r f_{ri} \right) = \frac{1}{Z_q} N q_i \exp \left( - \sum_{r=1}^{R} \lambda_r f_{ri} \right), \qquad i = 1, ..., s \tag{24}$$

or

$$p_i^* | q_i = \frac{n_i^*}{N} \approx q_i \exp \left( -\lambda_0 - \sum_{r=1}^{R} \lambda_r f_{ri} \right) = \frac{1}{Z_q} q_i \exp \left( - \sum_{r=1}^{R} \lambda_r f_{ri} \right), \qquad i = 1, ..., s \tag{25}$$

with

$$Z_q = e^{\lambda_0} = \sum_{i=1}^{s} q_i \exp \left( - \sum_{r=1}^{R} \lambda_r f_{ri} \right) \tag{26}$$

where $p_i$ is the proportion or probability of entities in each state $i$. Since they contain the *a priori* probabilities $q_i$, (24)-(25) can be termed the generalized *Maxwell-Boltzmann cross-distribution*, whilst $Z_q$ is the generalized *cross-partition function* and $\lambda_0 = \ln Z_q$ is

the generalized *Massieu function* (strictly speaking, its negative [10, 82]). The Lagrangian multipliers are obtained from the constraints C$r$ (17) and/or more readily from moment calculations (see section II C 6).

If $\mathbf{q} = \mathbf{u}$, (25) reduces to:

$$p_i^*|u \approx \frac{1}{Z} \exp\left(-\sum_{r=1}^{R} \lambda_r f_{ri}\right), \qquad i = 1, ..., s$$
$$Z = \sum_{i=1}^{s} \exp\left(-\sum_{r=1}^{R} \lambda_r f_{ri}\right) \tag{27}$$

This is the generalized *Maxwell-Boltzmann distribution* of statistical thermodynamics and information theory, and $Z$ is the generalized *partition function* [8, 10, 12]. Note (27) is obtained directly if either $\ln \mathbb{P}|\mathbf{u}$ (14) or $\ln \mathbb{W}$ (15) is used in the Lagrangian (21) instead of $\ln \mathbb{P}$.

In the information literature, it is customary to cast the analysis in terms of $p_i$ rather than $n_i$, thus in "$p_i$ form" [8, 9, 10, 12]. The constraints are:

$$\text{C0}: \qquad \sum_{i=1}^{s} p_i = 1 \tag{28}$$

$$\text{C1 to C}R: \qquad \sum_{i=1}^{s} p_i f_{ri} = \langle f_r \rangle, \qquad r = 1, ..., R \tag{29}$$

hence the Lagrangian (21) is:

$$\mathfrak{L} = \sum_{i=1}^{s} (p_i \ln N! - \ln[(p_i N)!] + p_i N \ln q_i)$$
$$-(\mu_0 - N)\left(\sum_{i=1}^{s} p_i - 1\right) - \sum_{r=1}^{R} \mu_r \left(\sum_{i=1}^{s} p_i f_{ri} - \langle f_r \rangle\right) \tag{30}$$

where $\mu_r$, $r = 0, ..., R$, are the new Lagrangian multipliers, and $(\mu_0 - N)$ is used for convenience. Taking the variation and applying the Stirling approximation gives:

$$p_i^*|q_i \approx q_i \exp\left(-\frac{\mu_0}{N} - \sum_{r=1}^{R} \frac{\mu_r}{N} f_{ri}\right) = \frac{1}{Z'_q} q_i \exp\left(-\sum_{r=1}^{R} \frac{\mu_r}{N} f_{ri}\right), \qquad i = 1, ..., s \tag{31}$$

with

$$Z'_q = e^{\mu_0/N} = \sum_{i=1}^{s} q_i \exp\left(-\sum_{r=1}^{R} \frac{\mu_r}{N} f_{ri}\right) \tag{32}$$

This is identical to (25)-(26), with $\lambda_r = \mu_r/N$, $r = 0, ..., R$ and $Z'_q = Z_q$. The Lagrangian multipliers are again obtained from the constraints (29).

It is worth commenting that if the leading $\ln N!$ term is not brought inside the summation in (20), but discarded - the approach of all previous workers - the resulting distribution $p_i^*$ contains an additional dependence on $N^{-1}$, which cancels out when forming the partition function $Z''_q = Ne^{\lambda_0}$. It therefore has no effect on traditional statistical mechanics. The distinction is, however, important in the development of exact statistical mechanics, as reported in Part II.

From the foregoing it is clear that the "most probable" probability distribution for a multinomial system, subject to arbitrary moment constraints, can be obtained *without making use of an entropy or cross-entropy function.* One can instead analyse a probabilistic system directly using first combinatorial principles. This aspect of entropy theory is not clearly spelt out in the information theory literature, with only a few exceptions (e.g. [17, 18, 47, 48, 49, 50]). The direct combinatorial approach is examined further in section III B, for systems not of multinomial character.

### *3. Definition of the Cross-Entropy (Directed Divergence) and Entropy*

Where do the cross-entropy and entropy functions come into the above analyses? Clearly, they are merely convenient mathematical tools to enable construction of the Lagrangian equation in $p_i$ form (30). In fact we can *define* the cross-entropy as "that function which, when inserted into the Lagrangian in place of $\ln \mathbb{P}$, and the extremum of the Lagrangian is obtained, yields the most probable cross-distribution of the system". The entropy may be similarly *defined* as "that function which, when inserted into the Lagrangian in place of $\ln \mathbb{P}|\mathbf{u}$ (or $\ln \mathbb{W}$), and the extremum of the Lagrangian is obtained, yields the most probable distribution of the system."

Consider $\ln \mathbb{P}$, expressed in $p_i$ form:

$$\ln \mathbb{P} = \sum_{i=1}^{s} (p_i \ln N! - \ln[(p_i N)!] + p_i N \ln q_i) \tag{33}$$

whence from the Stirling approximation (23) [50]:

$$\ln \mathbb{P} \approx \sum_{i=1}^{s} (p_i(N \ln N - N) - (p_i N \ln(p_i N) - p_i N) + p_i N \ln q_i)$$
$$= -N \sum_{i=1}^{s} p_i \ln \tfrac{p_i}{q_i} = -ND \tag{34}$$

Thus the cross-entropy or directed divergence $D$ (6) is simply the negative of the logarithm of the governing probability distribution, expressed per number of entities present [50]. Maximizing $\ln \mathbb{P}$ for a multinomial system subject to the Stirling limits is therefore equivalent to maximizing $-D$, or minimizing $D$, subject to the constraints on the system. (It does not matter whether we adopt a positive function, whose minimum yields the most probable realization, or its negative, whose maximum also yields this realization. By convention, the cross-entropy is taken here as a positive function to be minimized, although this choice is arbitrary.)

Similarly if we consider $\ln \mathbb{P}|\mathbf{u}$, from (28) and (34) the Stirling form is [3, 18, 50]:

$$\ln \mathbb{P}|\mathbf{u} \approx -N \sum_{i=1}^{s} p_i \ln s p_i = -N \ln s + NH \tag{35}$$

This is proportional to the Shannon entropy (5), shifted by a constant. Maximizing $\ln \mathbb{P}|\mathbf{u}$ subject to the Stirling limits and constraints is therefore equivalent to maximizing $H$, subject to the same constraints [50]. Indeed, from (15),

$$\ln \mathbb{W} \approx -N \sum_{i=1}^{s} p_i \ln p_i = NH \tag{36}$$

This definition of entropy for a multinomial system accords with the probabilistic expressions of Boltzmann and Shannon ((4)-(5)).

It is therefore seen that the Kullback-Leibler cross-entropy and Shannon entropy functions are simplified forms of the logarithm of the multinomial distribution (13), expressed per unit entity. The MinXEnt and MaxEnt principles therefore provide simplified methods to determine the most probable realization of a multinomial system, subject to its constraints. The cross-entropy is the more generic of the two functions, in that it contains the *a priori* probabilities $q_i$.

Of the three theoretical roots of the entropy and cross-entropy functions, the combinatorial approach is therefore the most intellectually satisfying in that it provides a direct answer to the question: what is entropy? There is no circular argument: entropy and cross-entropy

are firmly based on simple combinatorial principles. In consequence, there is no need to imbue either the MinXEnt or MaxEnt principles, or the cross-entropy or entropy functions themselves, with the kind of mystique with which they have been associated for well over a century. There is no mystery at all. In later sections, the foregoing analysis is generalized to any probabilistic system, irrespective of whether it is of multinomial character.

*4. Equivalence of Reference States*

It is necessary to be extremely careful about the definitions of the cross-entropy and entropy functions, given in section II C 3. To this end, note that obtaining the extremum of the Lagrangian ((21) or (30)) necessitates extremization, whether it contains $\ln \mathbb{P}$ or its substitute, $-D$ (or whether $\ln \mathbb{W}$ or $H$, if $\mathbf{q} = \mathbf{u}$). The relationship between these quantities is therefore:

$$d\left(-D(\mathbf{p}|\mathbf{q})\right) = \frac{1}{N} d\left(\ln \mathbb{P}\right) \tag{37}$$

(In the present analysis, $\lambda_0, ..., \lambda_R$ in the Lagrangian can be multiplied by any arbitrary positive constant $K$, and still give the same distribution, and so we could relax (37) further by extremizing the scaled negative cross-entropy $-KD$. This explains why we can use the scaled entropy $S = kH$ (4) throughout thermodynamics, without affecting any calculations.) Correspondence between the $i$th terms of $D$ and $\ln \mathbb{P}$ gives:

$$-\frac{\partial}{\partial p_i} D_i(p_i|q_i) dp_i = \frac{1}{N} \frac{\partial}{\partial p_i} \ln \mathbb{P}_i dp_i \qquad i = 1, ..., s \tag{38}$$

where

$$D(\mathbf{p}|\mathbf{q}) = \sum_{i=1}^{s} D_i(p_i|q_i), \qquad \ln \mathbb{P} = \sum_{i=1}^{s} \ln \mathbb{P}_i$$

Integration with respect to $p_i$ and summation gives:

$$-D(\mathbf{p}|\mathbf{q}) = \frac{1}{N} \sum_{i=1}^{s} \int \frac{d}{dp_i} \ln \mathbb{P}_i dp_i = \frac{\ln \mathbb{P}}{N} + C \tag{39}$$

where $C$ is a constant of integration. In consequence, the multinomial cross-entropy (6) and entropy (5) could have been given respectively as (or any multiple of):

$$D(\mathbf{p}|\mathbf{q}) = C + \sum_{i=1}^{s} p_i \ln \frac{p_i}{q_i} = \sum_{i=1}^{s} \left( C p_i + p_i \ln \frac{p_i}{q_i} \right) \tag{40}$$

$$H(\mathbf{p}) = -C - \sum_{i=1}^{s} p_i \ln p_i = -\sum_{i=1}^{s} (Cp_i + p_i \ln p_i) \tag{41}$$

However, the axiomatic definitions of these functions require that they obey the decisivity property (section II B), i.e. $D = H = 0$ when $\{p_i = 1, i = j;\ p_i = 0, i \neq j\}$, from which $C = 0$, producing the recognized forms of the above functions ((6), (5)). This causes the $N \ln s$ term to be dropped from the definition of $H$ (35). Note, however, that the choice of $C$ has no impact on the application of $D$ or $H$ to determine the most probable realization. (In other words, as is recognized throughout science and engineering, all zero reference or datum positions for the cross-entropy and entropy - and hence for information and energy - are mathematically equivalent.) This subtle point is examined further in Part II, in relation to exact thermodynamics.

*5. Ensemble Theory and Multicomponent Systems*

In its application to thermodynamics, one aspect of statistical mechanics has caused needless conceptual difficulty: the use of *ensembles* to represent particular types of systems [83]. Most common are the *microcanonical ensemble*, representing a closed system of fixed energy; the *canonical ensemble*, a closed system of fixed temperature; and the *grand canonical ensemble*, an open system of fixed temperature and mean composition. From the foregoing discussion, it is evident that an ensemble is simply *the set of all possible realizations - each weighted by its number of permutations (or for unequal $q_i$, by the probability of each permutation) - consistent with a particular system specification*; i.e. consistent with a specified governing probability distribution $\mathbb{P}$, total number of entities $N$ (which may include entities of different types), number of states $s$, and specified constraints $\langle f_r \rangle$ or their equivalent Lagrangian multipliers $\lambda_r, r = 1, ..., R$. An ensemble is therefore a *mental* construct, which does not require a physical manifestation.

As an example, consider a closed physical system in which the entities fluctuate between states (the *elemental chaos* of Planck [5]), such as gas molecules in a container. Such a system will migrate from one realization to another, and thence between different members of its ensemble (it will describe a trajectory in - for example - energetic, phase or system space). However, there is no need to require that the system *must* access every realization within a particular time frame, nor even that it should come arbitrary close to every realization;

the only requirement of probability theory is that each realization included in the ensemble be realizable, to the extent given by its assigned probability. As every gambler or insurance broker will testify, probabilities are not certainties. Unfortunately, a great deal of erroneous reasoning has been put forth on this topic, which still clouds our present-day understanding.

In contrast, consider a "multiple selection" system as defined in section II C 1, such as a set of throws of a coin or rolls of a die. In this case, the ensemble can only ever be a mental construct, representing the set of all possible outcomes. Once the "die is cast", the ensemble ceases to have any meaning, except as a reminder of what "might have been".

The microcanonical and canonical ensembles are both based on the multinomial distribution (13), with different interpretations. In the (generalized) microcanonical ensemble, $N$ represents the total number of non-interacting particles, each of which is deemed to possess its own "private" functions $f_{ri}$. The constraints $\langle f_r \rangle$ can therefore be kept constant. In contrast, in the (generalized) canonical ensemble, $N$ is now the number of separate systems (this is more clearly denoted $\mathbb{N}$; [46]), each of which contains a constant number of particles, all subject to baths of constant $\lambda_r, r = 1, ..., R$. By this device, the canonical ensemble can be used to examine systems containing interacting particles [5] or other coupling effects, thus in which the $f_{ri}$ (hence the $\langle f_r \rangle$) can be functions of the realization, even though the $\lambda_r$ are fixed (see [39, 83, 84, 85]). In other words, the canonical ensemble represents "the set of realizations of the set of realizations of interacting particles." This superset cannot readily be reduced to the lower (microcanonical) set unless the particles are non-interacting. Despite this distinction, by the use of baths of "generalized heat" (see section II C 6), the canonical ensemble is analysed by the same mathematical treatment as the microcanonical [8, 39].

The generalized grand canonical ensemble is normally taken to consist of $\mathbb{N}$ separate systems, in which there are $n_{\{N_l\},i}$ systems containing $N_l$ particles each of the $l$th type in the $i$th state, for $l = 1, ..., L$, where $L$ is the number of independent species. (For reactive species, it is necessary to define a minimum set of $L$ species, from which all other species can be formed by reaction [86].) Since the system is open, each $N_l$ is permitted to vary between zero and (effectively) infinity. Expressed in terms of $\mathbb{P}$ rather than $\mathbb{W}$, the governing distribution

[5] The precise definition of "interacting" remains open. Some workers prefer to qualify this statement, by considering only "weakly interacting" particles (e.g. [8, 42]) or those without "long-range interactions" (e.g. [25]).

is generally assumed to be "multiply multinomial" (c.f. [40, 43, 44, 45]):

$$\mathbb{P}_{GC} = \mathbb{N}! \prod_{N_1=0}^{\infty} ... \prod_{N_L=0}^{\infty} \prod_{i=1}^{s} \frac{q_{\{N_l\},i}^{\;\;n_{\{N_l\},i}}}{n_{\{N_l\},i}!} \tag{42}$$

where $q_{\{N_l\},i}$ is the *a priori* probability of a system which contains $N_l$ particles of each $l$th type in the $i$th state. This is normally subject to natural, moment and mean number of each type of entity constraints:

$$\mathrm{C0}: \qquad \sum_{N_1=0}^{\infty} ... \sum_{N_L=0}^{\infty} \sum_{i=1}^{s} n_{\{N_l\},i} = \mathbb{N} \tag{43}$$

$$\mathrm{C}r: \qquad \sum_{N_1=0}^{\infty} ... \sum_{N_L=0}^{\infty} \sum_{i=1}^{s} n_{\{N_l\},i} f_{ri} = \mathbb{N} \left\langle f_r \right\rangle, \qquad r = 1, ..., R \tag{44}$$

$$\mathrm{C}l: \qquad \sum_{N_1=0}^{\infty} ... \sum_{N_L=0}^{\infty} \sum_{i=1}^{s} n_{\{N_l\},i} N_l = \mathbb{N} \left\langle N_l \right\rangle, \qquad l = 1, ..., L \tag{45}$$

The combinatorial method ((39) and section II C 2) gives the Stirling-approximate cross-entropy and equilibrium distribution:

$$-D_{GC} = \frac{\ln \mathbb{P}}{\mathbb{N}} \approx -\sum_{N_1=0}^{\infty} ... \sum_{N_L=0}^{\infty} \sum_{i=1}^{s} p_{\{N_l\},i} \ln \frac{p_{\{N_l\},i}}{q_{\{N_l\},i}} \tag{46}$$

$$p^*_{\{N_l\},i} = \frac{1}{\Xi_q} q_{\{N_l\},i} \exp\left( -\sum_{r=1}^{R} \lambda_r f_{ri} - \sum_{l=1}^{L} \nu_l N_l \right), \qquad i = 1, ..., s \tag{47}$$

with

$$\Xi_q = \sum_{N_1=0}^{\infty} ... \sum_{N_L=0}^{\infty} \sum_{i=1}^{s} q_{\{N_l\},i} \exp\left( -\sum_{r=1}^{R} \lambda_r f_{ri} - \sum_{l=1}^{L} \nu_l N_l \right) \tag{48}$$

where $p_{\{N_l\},i} = n_{\{N_l\},i}/\mathbb{N}$; $\lambda_r$ and $\nu_l$ are Lagrangian multipliers; and $\Xi_q$ is the grand cross-partition function. The entropy forms follow. However, the cross-entropy will only be of Kullback-Leibler form if the governing distribution is multinomial (42). If $\mathbb{P}$ is of some other form, for example the product of independent distributions (extending [41]):

$$\mathbb{P}_{GC}{}' = \prod_{l=0}^{L} \mathbb{P}_l = \prod_{l=0}^{L} \prod_{N_l=0}^{\infty} \mathbb{N}_l! \prod_{i=1}^{s} \frac{q_{N_l,i}^{\;\;n_{N_l,i}}}{n_{N_l,i}!} \qquad \sum_{N_l=0}^{\infty} \sum_{i=1}^{s} n_{N_l,i} = \mathbb{N}_l \tag{49}$$

or if we possess some other knowledge (such as of $N_l$), then clearly the resulting multicomponent cross-entropy and entropy functions and the equilibrium distribution could be quite different. It is insufficient to simply assert (42) or (46); its adoption must be based on sound reasoning, and ultimately, be demonstrated by successful predictions.

*6. "Jaynes Relations" and Generalized Free Energy Function*

For completeness and to assist the analysis in Part II, the main implications of the foregoing analysis (section II C 2) are surveyed here, by synthesis and extension of previous treatments (primarily due to Jaynes [8, 18, 48]; see also [9, 10, 12]). Throughout the following (except where specified), $\lambda_0 = \ln Z_q$ is assumed to be a function of each $\lambda_r$; the $\lambda_r$ are mutually independent; each $f_{ri}$ is independent of $\lambda_r$; and each $\langle f_r \rangle$ is a function of $\lambda_r$ but not of the other multipliers $\lambda_m$, $m \neq r$. From $p_i^*$ (24)-(25) and the moment constraints (17) it can be shown that [8, 10, 12]:

$$-\frac{\partial \lambda_0}{\partial \lambda_r} = \langle f_r \rangle \tag{50}$$

The variance and covariances of $f_{ri}$, necessarily in the vicinity of equilibrium, are obtained by further differentiation [8, 9, 10, 12, 18]:

$$\frac{\partial^2 \lambda_0}{\partial \lambda_r^2} = \mathrm{var}(f_r) = \left\langle f_r^2 \right\rangle - \langle f_r \rangle^2 = -\frac{\partial \langle f_r \rangle}{\partial \lambda_r} \tag{51}$$

$$\frac{\partial^2 \lambda_0}{\partial \lambda_m \partial \lambda_r} = \mathrm{cov}(f_m, f_r) = \langle f_r f_m \rangle - \langle f_r \rangle \langle f_m \rangle = -\frac{\partial \langle f_r \rangle}{\partial \lambda_m} \tag{52}$$

where each $f_{ri}$ is independent of each $\lambda_m$. From (52), $\partial^2 \lambda_0 / \partial \lambda_m \partial \lambda_r = \partial^2 \lambda_0 / \partial \lambda_r \partial \lambda_m$, whence the coupling coefficients are equal:

$$\frac{\partial \langle f_r \rangle}{\partial \lambda_m} = \frac{\partial \langle f_m \rangle}{\partial \lambda_r} \tag{53}$$

Note (52) is a subset of a more general result [48]:

$$\mathrm{cov}(g, f_r) = \langle g f_r \rangle - \langle g \rangle \langle f_r \rangle = -\frac{\partial \langle g \rangle}{\partial \lambda_r} \tag{54}$$

where $\mathbf{g} = \{g_i\}$ is any function of the states $i = 1, ..., s$, in which each $g_i$ is independent of $\lambda_r$.

Using the Cauchy-Schwartz inequality $\langle a^2 \rangle \langle b^2 \rangle - \langle ab \rangle^2 \geq 0$ [87] with $a = f_r, b = 1$ gives $\mathrm{var}(f_r) = -\partial \langle f_r \rangle / \partial \lambda_r \geq 0$, whence $\partial \langle f_r \rangle / \partial \lambda_r \leq 0$ [11]. Accordingly, $\lambda_r$ decreases monotonically with increasing $\langle f_r \rangle$. No equivalent relation is available for the mixed derivatives $\partial \langle f_r \rangle / \partial \lambda_m$. Using the arguments of Kapur & Kesevan [12: sections 2.4.2; 4.3.2], we find that $\lambda_0$ is a convex function of $\lambda_r$, $r = 1, ..., R$.

It is also possible to consider $\lambda_0$ and each $f_{ri}$ (hence also $\langle f_r \rangle$) to be functions of parameters $\alpha_v$, $v = 1, ..., V$. By differentiation of the cross-partition function (26) [8, 18, 48], or more

directly by rearrangement of $p_i^*$ ((24)-(25)) and differentiation:

$$-\frac{\partial \lambda_0}{\partial \alpha_v} = \sum_{r=1}^{R} \lambda_r \left\langle \frac{\partial f_r}{\partial \alpha_v} \right\rangle, \qquad v = 1, ..., V \tag{55}$$

Alternatively, differentiation of (53) with respect to any continuous function $\alpha_v$ yields (necessarily in the vicinity of equilibrium, e.g. for a shifting equilibrium position):

$$\frac{\partial}{\partial \lambda_m} \left( \frac{\partial \langle f_r \rangle}{\partial \alpha_v} \right) = \frac{\partial}{\partial \lambda_r} \left( \frac{\partial \langle f_m \rangle}{\partial \alpha_v} \right) \tag{56}$$

Note (56) with $\alpha_v = t$ =time is a statement of Onsager's [88, 89] reciprocal relations. Various other higher derivative equations in $\lambda_r$ and/or $\alpha_v$ are given by Jaynes [48].

Similarly, considering $\lambda_0$ and $\lambda_r$ to be functions of $\beta_j$, $j = 1, ..., J$; or $\lambda_0$ alone as a function of $N$, $n_i^*$ or $p_i^*$, from (24)-(26):

$$-\frac{\partial \lambda_0}{\partial \beta_j} = \sum_{r=1}^{R} \frac{\partial \lambda_r}{\partial \beta_j} \langle f_r \rangle, \qquad j = 1, ..., J \tag{57}$$

$$\frac{\partial \lambda_0}{\partial N} = 0 \tag{58}$$

$$-\frac{\partial \lambda_0}{\partial n_i^*} = \frac{1}{n_i^*}, \qquad -\left\langle \frac{\partial \lambda_0}{\partial n_i^*} \right\rangle = \left\langle \frac{1}{n_i^*} \right\rangle = \frac{s}{N} \tag{59}$$

$$-\frac{\partial \lambda_0}{\partial p_i^*} = \frac{1}{p_i^*}, \qquad -\left\langle \frac{\partial \lambda_0}{\partial p_i^*} \right\rangle = \left\langle \frac{1}{p_i^*} \right\rangle = s \tag{60}$$

From (58), $\lambda_0$ (and thus $Z_q$) is independent of $N$ in the Stirling limit $N \to \infty$. From (59), $\langle \partial \lambda_0 / \partial n_i^* \rangle \to 0$ in the Stirling limit $n_i^* \to \infty$, hence $\lambda_0$ is independent of the mean degree of filling of each state.

Using $p_i^*$ ((24)-(25)), the constraints ((16)-(17) or (28)-(29)), the definitions of $H$, $D$ and $\mathbb{P}$ ((5)-(6),(39)) and the multiplier relations ((50)), the minimum cross-entropy or maximum entropy position is obtained as (c.f. [8, 10, 12]):

$$-D^* = H^* = \lambda_0 + \sum_{r=1}^{R} \lambda_r \langle f_r \rangle = \ln Z_q - \sum_{r=1}^{R} \lambda_r \frac{\partial \ln Z_q}{\partial \lambda_r} \tag{61}$$

with probability:

$$\mathbb{P}^* = A \exp(-N D^*) \tag{62}$$

where $A$ is a normalising constant (with $\mathbb{P}^* \leq 1$), and we recall that $H^*$ is obtained from $\ln \mathbb{P}|\mathbf{u}$ by dropping the $\ln s$ term (or directly from $\ln \mathbb{W}$) ((35)-(36)). Note (61) is one of the most important equations in equilibrium statistical mechanics - for example giving the

thermodynamic entropy and thence all thermodynamic functions in terms of the applicable partition function - whilst (62) encompasses Einstein's [90] definition of entropy. Note that the minimum cross-entropy and maximum entropy positions are of the same form, although $\mathbf{q}$ is implicit within $\lambda_0$ in $D^*$. By successive differentiation of (61) with respect to the moments - taking $\lambda_0$ to be independent of $\langle f_r \rangle$ - gives (c.f. [8, 12, 18, 48]):

$$-\frac{\partial D^*}{\partial \langle f_r \rangle} = \frac{\partial H^*}{\partial \langle f_r \rangle} = \lambda_r \tag{63}$$

$$-\frac{\partial^2 D^*}{\partial \langle f_m \rangle \partial \langle f_r \rangle} = \frac{\partial^2 H^*}{\partial \langle f_m \rangle \partial \langle f_r \rangle} = \frac{\partial \lambda_r}{\partial \langle f_m \rangle} = \frac{\partial \lambda_m}{\partial \langle f_r \rangle} \tag{64}$$

whilst differentiation with respect to $\lambda_r$ - now considering $\langle f_r \rangle$ to be a function of $\lambda_m, \forall m$ - and use of (53) gives the Euler relation (c.f. [91]):

$$-\frac{\partial D^*}{\partial \lambda_r} = \frac{\partial H^*}{\partial \lambda_r} = \sum_{m=1}^{M} \lambda_m \frac{\partial \langle f_m \rangle}{\partial \lambda_r} = \sum_{m=1}^{M} \lambda_m \frac{\partial \langle f_r \rangle}{\partial \lambda_m} \tag{65}$$

where $M$ and $R$ are numerically equal. From (63), using the same arguments as Kapur & Kesevan [12: sections 2.4.4; 4.3.2], we see that $D^*$ (or $H^*$) is a convex (concave) function of the $\langle f_r \rangle$'s. A multinomial system subject to the Stirling approximation therefore has a single, unique equilibrium position with respect to its moment constraints.

The variation in $D^*$ or $H^*$ due to variations in $\lambda_0$, $\lambda_r$ and $\langle f_r \rangle$ (and also $N$) is (c.f. [8, 10, 18, 48]):

$$-dD^* = dH^* = \sum_{r=1}^{R} \lambda_r (d\langle f_r \rangle - \langle df_r \rangle) = \sum_{r=1}^{R} \lambda_r dQ_r \tag{66}$$

where we can interpret $d\langle f_r \rangle = dU_r$, $\langle df_r \rangle = \sum_{i=1}^{s} p_i df_{ri} = dW_r$ and $d\langle f_r \rangle - \langle df_r \rangle = \sum_{i=1}^{s} f_{ri} dp_i = dQ_r$ respectively as changes in the $r$th type of "energy", "generalized work" on the system and "generalized heat" delivered to the system , whence (as defined here) $dU_r = dQ_r + dW_r$. Note that in the above derivation, the variations in $\lambda_r$ cancel out ([18, 48]), hence (66) encompasses conditions of either constant or variable $\lambda_r$. Note (66) is a superset of the Clausius relation (1), and so for each type of "generalized heat" there exists a conjugate integrating factor $\lambda_r$. As with the Clausius relation, the $\lambda_r$ are properties of the system of interest (i.e. the one into which positive generalized heat is delivered).

Equation (66) applies to a reversible process, i.e. to an incremental change in the equilibrium position. If we also include spontaneous irreversible processes (involving a system not necessarily at equilibrium), for which the cross-entropy can decrease (or entropy can

increase) without generalized heat input, we see that:

$$-dD = dH \geq \sum_{r=1}^{R} \lambda_r dQ_r \tag{67}$$

This is a superset of the Clausius inequality (2). Note (67) can be rearranged, in the manner of Gibbs [83, 86], to give the differential form of a generic dimensionless free energy function $\Phi$, here termed the *free information*:

$$d\Phi = \left\{ \begin{array}{c} dD + \sum\limits_{r=1}^{R} \lambda_r dQ_r \\ -dH + \sum\limits_{r=1}^{R} \lambda_r dQ_r \end{array} \right\} \leq 0 \tag{68}$$

(whence $d\Phi^* = 0$ at a fixed equilibrium position), where the upper form incorporates the *a priori* probabilities $\mathbf{q}$. Now from (61):

$$-dD^* = dH^* = d\lambda_0 + \sum_{r=1}^{R} d\lambda_r \langle f_r \rangle + \sum_{r=1}^{R} \lambda_r d\langle f_r \rangle \tag{69}$$

so if we set $dD = dD^* + dD^{irrev}$ and $dH = dH^* + dH^{irrev}$ (with $dD^{irrev} \leq 0$ and $dH^{irrev} \geq 0$), where superscript *irrev* denotes the irreversible component, then from (68)-(69):

$$d\Phi = \left\{ \begin{array}{c} -d\lambda_0 - \sum\limits_{r=1}^{R} d\lambda_r \langle f_r \rangle + dD^{irrev} - \sum\limits_{r=1}^{R} \lambda_r dW_r \\ -d\lambda_0 - \sum\limits_{r=1}^{R} d\lambda_r \langle f_r \rangle - dH^{irrev} - \sum\limits_{r=1}^{R} \lambda_r dW_r \end{array} \right\} \leq 0 \tag{70}$$

If - and only if - there is no change in $\lambda_r$ (i.e. no change in any contacting bath; see also (74) below), no reversible generalized work on the system (apart from that already included in the constraints) and no irreversible process, then:

$$d\Phi^* = -d\lambda_0 = -d \ln Z_q \tag{71}$$

where $Z_q$ is the applicable cross-partition or partition function ((26) or (27)). Similarly, from (70), if there is no change in $\lambda_0$ or $\lambda_r$ and no irreversible process:

$$d\Phi = -\sum_{r=1}^{R} \lambda_r dW_r \leqslant 0 \tag{72}$$

$\Phi$ therefore indicates the maximum available weighted generalized work per entity which can be obtained from a system.

Integration of (68) gives the state function:

$$\Phi = \begin{cases} D + \sum_{r=1}^{R} \lambda_r Q_r \\ -H + \sum_{r=1}^{R} \lambda_r Q_r \end{cases} \tag{73}$$

where $Q_r = \int dQ_r = \int d\langle f_r \rangle - \int dW_r$ defines each absolute generalized heat [6] . Comparing its differential with (68) gives:

$$\sum_{r=1}^{R} Q_r d\lambda_r = 0 \tag{74}$$

This is a superset of the Gibbs-Duhem equation [86]. For a system containing separate coexistent *phases*, or bodies which differ in composition or state (as defined by Gibbs [86]), there will be one such equation for each phase. For $L$ independent constituents, $\mathfrak{r} = R - L$ other constraints (not including the $L$ constituents) and $\mathfrak{p}$ phases, (74) thus yields a generalized Gibbs' phase rule for the number of degrees of freedom of a system (c.f. [10, 42, 86]):

$$f = L + \mathfrak{r} - \mathfrak{p} = R - \mathfrak{p} \tag{75}$$

In other words, the system will be fully determined by $R - \mathfrak{p}$ independent parameters, from the set of $R$ constraints or (more commonly) their corresponding Lagrangian multipliers.

Equations (61), (68) and (70)-(75) form the basis of present-day thermodynamics. For energetic systems, $d\Phi$ is normally divided by the energetic multiplier $\lambda_1 = 1/kT$; e.g. for an energetic system which can exchange heat with its surroundings, but not work or mass, at constant volume, $dQ_1 = dU$, $dS = kdH$, $dA = kTd\Phi = dU - TdS \leq 0$ and $dA^* = -kTd\ln Z$, where $U$ is the mean internal energy per entity, $A$ is the Helmholtz free energy per entity and $Z$ is the microcanonical or canonical partition function [7] . For a grand canonical system with $L$ independent constituents which can exchange heat and mass with its surroundings, but not work except for $PV$-work, at constant pressure, $dQ_1 = dU$, $\lambda_1 = 1/kT$, $dQ_2 = dV$, $\lambda_2 = P/kT$, $dQ_{2+l} = dm_l$, $\lambda_{2+l} = -\mu_l/kT = -\ln \alpha_l$, $dG = kTd\Phi = dU - TdS + PdV -$

[6] In thermodynamic systems, this is generally approximated as $Q_r \approx \langle f_r \rangle$, i.e. assuming each generalized work term is zero, except for the energy constraint, where the actual heat $Q = \int dQ = \int TdS = TS$ at constant $T$ is used.

[7] The extensive thermodynamic variables (e.g. $U, S, V, m_l, A, G$) are all mean quantities, expressed in relevant units per entity. In a microcanonical ensemble, they represent mean values per particle. The total values are calculated by multiplication by $N$ (the form of (68) remains the same). In a canonical ensemble, each extensive variable represents the "ensemble mean" or "mean of the total values".

$\sum_l \mu_l dm_l \leqslant 0$, $dG^* = -kTd\ln\Xi$ and $f = L + 2 - \mathfrak{p}$, where $P$ is pressure, $V$ is mean volume per entity, $\mu_l$ is the chemical potential and $\alpha_l$ is the "absolute" (unscaled) chemical activity of the $l$th constituent, $m_l$ is the mean number of entities of $l$th type per entity, $G$ is the Gibbs free energy per entity and $\Xi$ is the grand canonical partition function. The *essergy* $Y = kT_0\Phi = E + T_0S + P_0V - \sum_l \mu_{l0}m_l$ is a scaled $\Phi$ of a system with total internal energy $E$, in contact with a bath of reference temperature $T_0$, pressure $P_0$ and chemical potentials $\mu_0$ [92]. Essergy is thus an extended free energy with reference to the bath (e.g. the external environment), not to the system. The *exergy* is the difference between the essergy of a system (by early authors, with the chemical potential terms omitted), and of the same system in equilibrium with the bath (e.g. [92, 93, 94, 95, 96, 97, 98, 99]). Exergy therefore represents the maximum work deliverable to the environment, by allowing a system to reach equilibrium with that environment. The statistical *extropy* [100, 101, 102] is a modified free information defined with respect to the reservoirs - with all generalized work terms set to zero (i.e. $Q_r \approx \langle f_r \rangle$) - less the modified free information at equilibrium. Exergy and extropy have been used as measures of environmental impact, i.e. as quantitative tools within and/or complementary to the framework of environmental life cycle assessment [98, 99, 100, 101].

Notwithstanding the historical development of this field, it must be emphasized that the use of $\Phi$ is not restricted to thermodynamic systems. Just as with the information entropy, we can define the free information of any probabilistic system - for example in communications, transport, urban planning, biology, geography, social science, politics, economics, linguistics, image analysis or any other field - and use it to examine its (probabilistic) stability. The entire armoury of state functions, cyclic integrals, Gibbs-Duhem and phase relations, Maxwell-like relations and Jaynes relations - currently considered the exclusive domain of thermodynamics - can then be brought to bear to the analysis of such systems [8].

### 7. *"Fluctuations" and Entropy Concentration Theorem*

Although the minimum cross-entropy or maximum entropy distribution is the "most probable", it cannot be *a priori* assumed to be the exclusive outcome. The sharpness of the predicted distribution has historically been examined by two methods: the fluctuation

[8] The possibility of discrete phases within a variety of systems, such as social, political or economic systems - and the "precipitation" and "melting" of such phases - is intriguing, and demands further investigation.

criterion of Gibbs [83] and Einstein [90], and the entropy concentration theorem of Jaynes [47, 48, 103, 104], in part foreshadowed by Boltzmann [105] and Einstein [106].

The first method examines the coefficient of variation $\delta$ of each constraining variable (or its square), commonly termed its "fluctuation" [9] . For a microcanonical system, this can be written as (c.f. [83, 90]):

$$\delta(Nf_r) = \frac{\sqrt{\mathrm{var}(Nf_r)}}{\langle Nf_r \rangle} = \frac{\sqrt{N\left[\langle f_r^2 \rangle - \langle f_r \rangle^2\right]}}{\langle Nf_r \rangle} \tag{76}$$

where we are careful with notation to consider the variability about the total extensive quantity $\langle Nf_r \rangle$ for a system of $N$ entities, not the variability of the fixed quantity per entity $\langle f_r \rangle$. (Of course, $\delta$ does not capture the full picture of the distribution of $N\{f_{ri}\}$, e.g. the skewness, kurtosis, etc, for which higher order moments must be considered.) The criterion for sharpness is normally stated as $\delta \ll 1$ [18, 90]. From (51) and (76):

$$\delta(Nf_r) = \frac{1}{\sqrt{N}} \sqrt{-\frac{1}{\langle f_r \rangle^2} \frac{\partial \langle f_r \rangle}{\partial \lambda_r}} \tag{77}$$

The term inside the second square root is positive, and in many cases of order unity, whereupon $\delta(Nf_r) \approx N^{-1/2} \to 0$ in the Stirling limit $N \to \infty$. For example, for a microcanonical system with $f_{1i} = \varepsilon_i$, $\langle f_1 \rangle = \langle \varepsilon \rangle = U$, $\lambda_1 = 1/kT$, containing an ideal monatomic non-interacting gas with $U = \frac{3}{2}kT$ and $C_v = \partial U/\partial T = \frac{3}{2}k$, where $C_v$ is the isovolumetric heat capacity per entity, we obtain $\delta(N\varepsilon) = (\frac{3}{2}N)^{-1/2} \approx N^{-1/2} \to 0$ (e.g. [38, 39, 40, 41, 42, 43, 45]) [10] . Although this result is not general (e.g. in the vicinity of phase changes [45]) it applies to many physical phenomena, producing what is widely regarded as the overwhelming precision of thermodynamics. If valid, the "$N^{-1/2}$ rule" applies only as $N \to \infty$; for small $N$, a second effect must also be considered (see Part II).

For the canonical and other ensembles, the variability of the (superset) $\{f_{ri}\}$ within *each* ensemble member is examined by (see above references):

$$\delta(f_r) = \frac{\sqrt{\mathrm{var}(f_r)}}{\langle f_r \rangle} = \frac{\sqrt{\left[\langle f_r^2 \rangle - \langle f_r \rangle^2\right]}}{\langle f_r \rangle} \tag{78}$$

[9] The term "fluctuation" is unfortunate, since it implies rapid change about the mean, which has little to do with the equilibrium position but depends on the system dynamics. $\delta(Nf_r)$ is simply a measure of the "variability" or "spread" of the equilibrium filling of $N\{f_{ri}\}$.

[10] All the listed authors consider $\delta(E)$ for a canonical ensemble, where $\langle E \rangle$ is the "mean of the total energies", but then take $\langle E \rangle = N \langle \varepsilon \rangle = \frac{3}{2}NkT$ for $N$ non-interacting particles - thus assuming the system is microcanonical - giving the same result.

whence from (50)-(51) and (71):

$$\delta(f_r) = \frac{1}{\langle f_r \rangle}\sqrt{-\frac{\partial \langle f_r \rangle}{\partial \lambda_r}} = \frac{1}{\langle f_r \rangle}\sqrt{\frac{\partial^2 \lambda_0}{\partial \lambda_r^2}} = \frac{1}{\langle f_r \rangle}\sqrt{-\frac{\partial^2 \Phi^*}{\partial \lambda_r^2}} \tag{79}$$

Whether or not this vanishes as $N \rightarrow \infty$ depends on the physical variable $r$ and the importance of interactions ([38, 40, 41, 75]; c.f. previous footnote ). The variability of $\{f_{ri}\}$ for the *total* ensemble can be examined using $\delta(\mathbb{N}f_r)$, where $\mathbb{N}$ is the number of ensemble members, giving a relation analogous to (77). It is commonly asserted that $\mathbb{N} \rightarrow \infty$ (e.g. [39]), an assumption scrutinized in more detail in Part II. If correct, the total ensemble will be heavily concentrated at its ensemble means $\langle f_r \rangle, \forall r$.

Jaynes' [47, 103, 104] entropy concentration theorem considers the relative importance of the equilibrium probability distribution $\mathbf{p}^* = \{p_i^*\}$ and some other distribution $\mathbf{p}' = \{p_i'\}$. From (34) or (62), the ratio of the probability of occurrence of $\mathbf{p}^*$ to that of $\mathbf{p}'$ is:

$$\frac{\mathbb{P}^*}{\mathbb{P}'} = \exp[N(-D^* + D')] \tag{80}$$

where $\mathbb{P}^*$, $\mathbb{P}'$ are the governing probability distributions and $D^*$, $D'$ are the cross-entropies corresponding respectively to $\mathbf{p}^*$ and $\mathbf{p}'$. This was originally formulated as the ratio of the number of ways in which $\mathbf{p}^*$ and $\mathbf{p}'$ can be realized [47, 106]:

$$\frac{\mathbb{W}^*}{\mathbb{W}'} = \exp[N(H^* - H')] \tag{81}$$

where $\mathbb{W}^*$, $\mathbb{W}'$ are the weights and $H^*$, $H'$ are the entropies corresponding to $\mathbf{p}^*$ and $\mathbf{p}'$. As shown by Jaynes [47, 103, 104], for $N \rightarrow 1000$ even a small difference in $H$ gives an enormous ratio, revealing the combinatorial dominance of the maximum entropy position.

Assuming $\mathbf{p}^*$, $\mathbf{p}'$ satisfy the constraints ((28)-(29)), and taking the Stirling limits $N \rightarrow \infty$ and $n_i \rightarrow \infty$, an analysis similar to Kapur & Kesavan [12: section 2.4.6] yields:

$$-D^* + D' = H^* - H' = \sum_{i=1}^{s} p_i' \ln\left(\frac{p_i'}{p_i^*}\right) \tag{82}$$

i.e. simply the directed divergence of $\mathbf{p}'$ from $\mathbf{p}^*$, from which $\mathbf{q}$ vanish (being incorporated into $\mathbf{p}^*$). Note (80)-(81) then give:

$$\frac{\mathbb{P}^*}{\mathbb{P}'} = \frac{\mathbb{W}^*}{\mathbb{W}'} = \exp\left\{N \sum_{i=1}^{s} p_i' \ln\left(\frac{p_i'}{p_i^*}\right)\right\} \tag{83}$$

If we now put $p'_i = p^*_i(1 + \varepsilon_i)$, take a series expansion of $\ln p'_i$ about $\varepsilon_i = 0$, and discard all polynomial terms higher than $\varepsilon_i^2$, it is shown by Kapur & Kesavan [12: section 2.4.7] that (a quite different derivation is given by Jaynes [104]):

$$-D^* + D' = H^* - H' \approx \frac{1}{2}\sum_{i=1}^{s}\frac{(p'_i - p^*_i)^2}{p^*_i} = \frac{1}{2N}\sum_{i=1}^{s}\frac{(n'_i - n^*_i)^2}{n^*_i} = \frac{1}{2N}\chi^2 \tag{84}$$

where $n'_i = p'_i N$ is the number of entities in state $i$ due to $\mathbf{p}'$; $n^*_i = p^*_i N$ is the expected number of entities in state $i$; and we recognize $\chi^2$ as the chi-squared distribution of statistics [20, 21, 107, 108]. In other words, we can determine the "goodness of fit" of a distribution $\mathbf{p}'$ - or of some function $F(p)$ which generates $\mathbf{p}'$ - to a multinomial system, by comparing the calculated $\chi^2$ to the table value $\chi^2(\nu, 1-\alpha)$, where $\nu = s - R - 1$ is the number of degrees of freedom and $\alpha$ is the significance level (upper tail or rejection area) [104].

As is well known [108, 109] and dramatically illustrated by Jaynes [48: chap 9], the $\chi^2$ statistic is an unreliable test for goodness of fit, being highly (and erroneously) sensitive to the occurrence of unlikely events. There is no need to conduct the simplification of (84); instead, from (82):

$$-D^* + D' = H^* - H' = \frac{1}{N}\sum_{i=1}^{s} n'_i \ln\left(\frac{n'_i}{n^*_i}\right) = \frac{\eta}{N} \tag{85}$$

where $\eta$ is the correct test statistic for the goodness of fit of $\mathbf{p}'$ or its generator $F(p)$ to a multinomial system, subject to the Stirling limits. ($\eta$ is given by Jaynes [48: section 9.11.1] in the form $\psi = 10\eta/\ln(10)$, using an obscure decibel notation.) The calculated $\eta$ can be compared to the "table value" $\eta(\nu, 1-\alpha)$; alternatively, two distributions $\mathbf{p}'$ and $\mathbf{p}''$ can be ranked by comparing their corresponding $\eta'$ and $\eta''$. Note (83) and (85) finally give:

$$\frac{\mathbb{P}^*}{\mathbb{P}'} = \frac{\mathbb{W}^*}{\mathbb{W}'} = \exp(\eta) \tag{86}$$

The exact form of $\eta$, which does not depend on the Stirling approximation, is examined in Part II.

## III. APPLICABILITY OF MULTINOMIAL STATISTICS

### A. The "Multinomial Family"

Why have the Shannon information entropy and Kullback-Leibler cross-entropy proved to be of such utility, in an extremely wide range of disciplines? The answer lies in the

fact that an extraordinarily large number of probability functions $p_{i,...}$ or $p(x,...)$ of an observable, encompassing a wide range of statistical problems, can be obtained from the Stirling approximation to the multinomial distribution as special or limiting cases. For example, in discrete statistics, the uniform, geometric, generalized geometric, power-function, Riemann zeta function, Poisson, binomial, negative binomial, generalized negative binomial and various Lagrangian distributions (and many others) have been obtained from the Shannon entropy subject to various constraints [12, 14]. Similarly, in continuous statistics, the uniform, normal (Gaussian), Laplace, generalized Cauchy, generalized logistic, generalized extreme value, exponential, Pareto, gamma, beta (of first or second kind), generalized Weibull, lognormal, Poisson, power-function and many new distributions, and various multivariate forms, can be obtained from the continuous form of the Shannon entropy subject to various constraints [12, 14]. Many additional distributions can be obtained from the Kullback-Leibler cross-entropy in discrete or continuous form, subject to various *a priori* distributions and constraints [12]. All these functions therefore constitute particular examples of multinomial statistics, and collectively form the *multinomial family* of statistical distributions. The broad applicability of the multinomial distribution, produced by the (fascinating) isomorphism of many probabilistic problems - such as of the "balls-in-boxes" and "multiple selection" systems described in section II C 1 - is responsible for the wide utility of the Kullback-Leibler cross-entropy and Shannon entropy functions.

### B. Non-Multinomial Statistics

Notwithstanding the success of multinomial statistics, it is important to emphasize that a number of statistical functions are incompatible with the Shannon entropy and/or Kullback-Leibler cross-entropy, and are therefore not of multinomial character. Several of these (e.g. Bose-Einstein, Fermi-Dirac, Rényi, Tsallis and Kaniadakis entropies) reduce to the Shannon entropy as a limiting case [22, 24, 25, 26, 27, 38, 41, 50]; such systems may therefore be approximated by multinomial statistics only when these limiting conditions are attained. A more comprehensive analysis of non-multinomial statistics must be deferred to a later work; however, their importance is here noted.

From the preceding analysis, it is clear that the definition of entropy (3) promulgated by Boltzmann [3] and Planck [4, 5] is correct irrespective of whether the distribution is

of multinomial character. A more comprehensive version, in which $\mathbb{P}$ now represents the governing probability distribution of *any* type and not only the multinomial distribution, is given in (39); this is a *generalized combinatorial cross-entropy*. From this a *generalized combinatorial entropy* is:

$$H(\mathbf{p}) = K\left(\frac{\ln \mathbb{P}|\mathbf{u}}{N} + C\right) = K\left(\frac{\ln \mathbb{W}}{N} + C'\right) \tag{87}$$

where $C$, $C'$ and $K$ are arbitrary constants. (Note that the Boltzmann [3] - Planck [4] formula (3) is often misleadingly quoted as $S = k \ln \mathbb{W}$; this is correct only if $S$ refers to the *total* entropy of the system, not the entropy per unit entity.) Indeed, it is not necessary to use a logarithmic transformation; for some distributions, some other metric $\phi(\mathbb{P}, N)$ may be more convenient, giving the even more generalized definitions:

$$-D_{gen}(\mathbf{p}, ...|\mathbf{q}, N...) = \kappa\left(\phi(\mathbb{P}, N, ...) + C\right) \tag{88}$$

$$H_{gen}(\mathbf{p}, ...|N, ...) = \kappa\left(\phi(\mathbb{P}|\mathbf{u}, N, ...) + C\right) = \kappa\left(\phi(\mathbb{W}, N, ...) + C'\right) \tag{89}$$

with the only condition on $\phi$ being:

$$\text{extr}\left[\phi(\mathbb{P}_{gen}, N, ...)\right] = \max\left[\mathbb{P}_{gen}\right] \tag{90}$$

where again $C$, $C'$ and $\kappa$ are arbitrary, whilst "..." allows for the presence of other variables [11] . Clearly, the information entropy (5) given by Shannon [6] - although derived from sound axiomatic postulates - is strictly valid only for multinomial systems subject to the Stirling approximation. This may be appropriate for communication signals of infinite length, but is surely insufficient to underpin the vast field of information theory in general.

Furthermore, Kapur, Kesevan and co-workers [11, 12, 53, 54] describe various *inverse methods* in maximum entropy theory, in which one works backwards from a hypothesized or observed probability distribution ($\mathbf{p}$), *a priori* distribution ($\mathbf{q}$) and constraints (C0-C$R$), to obtain the measure of cross-entropy or entropy applicable to the process. Using (88) or (89), such inverse methods can then be used to determine the governing probability distribution $\mathbb{P}$ of the process. Alternatively, one can work "sideways" from the observed, *a priori* and governing distributions ($\mathbf{p}$, $\mathbf{q}$ and $\mathbb{P}$) to determine the constraints (c.f. [54, 55, 56]).

[11] The recent derivation of the Tsallis [24] entropy by Suyari and co-workers [110, 111, 112, 113] using a transformation of the form $\phi = \ln_{2-q}(\mathbb{W}_{2-q})$, where $\ln_q$ is the $q$-logarithmic function and $\mathbb{W}_q$ is a $q$-multinomial coefficient, provides a fascinating example of an alternative metric.

Such methods offer powerful extensions to present-day information theory, to elucidate the fundamental probabilistic basis or constraints of a given statistical phenomenon.

As suggested by Topsøe [57, 58], one may also use *game theory* to develop the cross-entropy and entropy concepts. In this approach, a game is played between Player I ("Nature") and Player II ("a physicist"). Player II aims for low complexity, knowing the total set of available probability distributions $S_{II}$ chosen from an alphabet $\mathbb{A}$, whilst Player I aims for high complexity, knowing the subset $S_I$. Entropy arises from the choice of a complexity measure $\varphi(\mathbf{p}|\mathbf{q})$ of minimum complexity needed for Player II to determine $S_I$, i.e.:

$$H_\varphi(\mathbf{p}) = \inf_{\mathbf{q}\in S_{II}} \varphi(\mathbf{p}|\mathbf{q}) \tag{91}$$

where $\mathbf{p} \in S_I$ and $\mathbf{q} \in S_{II}$. The cross-entropy or divergence is given by the actual complexity minus minimal complexity:

$$D_\varphi(\mathbf{p}|\mathbf{q}) = \varphi(\mathbf{p}|\mathbf{q}) - H_\varphi(\mathbf{p}) \tag{92}$$

These concepts and further arguments are used to derive the Maxwell-Boltzmann, Tsallis and Kaniadakis entropies and cross-entropies, based on different generating functions [57, 58]. Clearly, the game-theoretic basis of entropy is deeply related to both Jaynes' MaxEnt principle and the combinatorial approach described herein, in a manner which deserves further examination. At very least, game theory with different strategies offers an alternative means to generate new cross-entropy or entropy functions, for which the governing distributions $\mathbb{P}$ can be identified using (88) or (89). It may also provide the means to generate entropy functions for which $\mathbb{P}$ is not readily expressed in closed mathematical form.

As a final comment, Jaynes in his many works expounds the "Bayesian" or "subjective" view of probabilities, which represent assignments of one's belief based on the available information, and argues against the "frequentist" or "objective" view in which probabilities are interpreted strictly as frequency assignments [8, 47, 114, 115]. Separately, Jaynes demonstrates the equivalence of MaxEnt based on the Shannon entropy, and combinatorial analysis using the multinomial weight [18, 47]. At this point, however, he considers the combinatorial approach to represent a frequency interpretation, stating [47, 48]: "the *probability* distribution which maximizes the entropy is numerically identical with the *frequency* distribution which can be realized in the greatest number of ways" [his emphasis]. This identification of the combinatorial approach with the frequentist view is unfortunate; in fact, by applying MaxEnt based on the Shannon entropy, one *assumes* (implicitly) that the phenomenon

being examined follows the multinomial distribution, and one uses one's prior knowledge to infer (hypothesize) the available states $i$ (for a parallel discussion, see Bhandari [116]). The calculated probability distribution $p_i^*$ is therefore valid only in the "subjective" sense (i.e. exists only as an inference of the observer) until verified by experiment. Even if so "verified", there will always be room for doubt over its validity. The broader Jaynesian program of maximum entropy analysis as a method of statistical inference is therefore untouched (in fact, enhanced) by the present analysis [12] .

Indeed, the definitions of cross-entropy and entropy given here ((88)-(89)) fit seamlessly into a more comprehensive Bayesian inferential framework (c.f. [72, 73, 115]), for probabilistic phenomena more complicated than those considered here. In such cases, $\mathbf{q}$ represents the "Bayesian prior distribution", "Jeffrey's uninformative prior" [117, 118] or "Jaynes' measure distribution" [18, 47], whilst $\mathbb{P}$ represents one's postulated understanding of the probabilistic structure of the phenomenon at hand.

The analysis to this point has followed a long path, only to arrive more or less at its starting point: the statistical entropy of Boltzmann and Planck (although the idea is taken somewhat further than they had imagined). The fact that this discussion is still necessary in the 21st century reflects the great gulf between present-day statistical mechanics and thermodynamics - still taught much as they were 50 or even 100 years ago - and the more recent but surprisingly narrow field of information theory initiated by Shannon [6]. The gulf persists despite the efforts of Bose, Einstein, Fermi and Dirac, amongst others, in statistical mechanics, and of Jaynes, Tribus, Kapur, Kesavan and many others in information theory and maximum entropy methods. The two fields are, in fact, one. Appreciation of this fact (by both sides) would permit the development of a much broader discipline of "combinatorial information theory" than at present, applicable to many different types of problems, including those examined in Part II.

## IV. CONCLUSIONS

In a detailed review and synthesis, the three main theoretical bases of the concepts of entropy and cross-entropy - information-theoretic, axiomatic and combinatorial - are crit-

[12] Jaynes appears to reach essentially this conclusion in his final work [48: chaps. 9 and 11; especially section 11.4].

ically examined. It is shown that the combinatorial basis, as promulgated by Boltzmann and Planck, is the most fundamental (most primitive) basis of these concepts. Not only does it provide (i) a derivation of the Kullback-Leibler cross-entropy and Shannon entropy functions, as simplified forms of the multinomial distribution subject to the Stirling approximation; the combinatorial approach also yields (ii) an explanation for the need to maximize entropy (or minimize cross-entropy) to find the most probable realization; and (iii) the means to derive entropy and cross-entropy functions for systems which do not satisfy the multinomial distribution, i.e. which fall outside the domain of the Kullback-Leibler and Shannon measures. The information-theoretic and axiomatic bases of cross-entropy and entropy - whilst of tremendous importance and utility - are therefore seen as secondary viewpoints, which lack the breadth of the combinatorial approach. The view of Shannon, Jaynes and their followers - in which the Shannon entropy or Kullback-Leibler cross-entropy is taken as the starting point and universal tool for analysis - is not seen as incorrect, but simply incomplete. On the other hand, the view of many scientists - who consider statistical mechanics to be a branch of classical mechanics or quantum physics, rather than a method for statistical inference in any field - is also incomplete. Appreciation of this reasoning would permit development of a powerful body of "combinatorial information theory", as a means for statistical inference in all fields (inside and outside science).

For completeness, the essential features of Jaynes' analysis of entropy and cross-entropy - reinterpreted in light of the combinatorial approach - are outlined, including derivation of probability distributions, Jaynes relations, a generalized free energy (or "free information") function, Gibbs-Duhem relation, phase rule, fluctuation theory, entropy concentration theorem, and generalized definitions of entropy and cross-entropy. The analysis is shown to be embedded within a Bayesian framework of statistical inference.

**Acknowledgments**

This work began in 2002, and was in part completed during sabbatical leave in 2003 at Clarkson University, New York; McGill University, Quebec; Rice University, Texas and Colorado School of Mines, Colorado, supported by The University of New South Wales and the Australian-American Fulbright Foundation. The work benefited from valuable discussions with participants at the 2005 NEXT Sigma Phi conference, Kolymbari, Crete, Greece.

**References**


[1] Clausius, R. (1865) Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie (On different forms of the fundamental equations of the mechanical theory of heat and their convenience for application), *Poggendorfs Annalen* 125: 335-400; *also in* Clausius, R. (1867) *Abhandlungen über die Mechanische Wärmetheorie (Treatise on the Mechanical Theory of Heat)*, F. Vieweg, Braunschweig, 2(IX): 1-44; English transl.: Lindsay, R.B. (1976), *in* Kestin, J. (ed.) (1976) *The Second Law of Thermodynamics*, Dowden, Hutchinson & Ross, Inc., PA, 162-193.

[2] Clausius, R. (1876) *Die mechanische Wrmetheorie (The Mechanical Theory of Heat)*, F. Vieweg, Braunschweig; English transl.: Browne, W.R. (1879), Macmillan & Co., London.

[3] Boltzmann, L. (1877), Über die Beziehung zwischen dem zweiten Hauptsatze dewr mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht (On the relation between the second law of the mechanical theory of heat and the probability calculus with respect to the theorems of thermal equilibrium), *Wien. Ber.*, 76: 373-435, English transl., Le Roux, J. (2002), 1-63, *http://www.essi.fr/l̃eroux/*.

[4] Planck, M. (1901), Über das gesetz der Energieverteilung im Normalspektrum (On the law of distribution of energy in the normal spectrum), *Annalen der Physik* 4: 553-563.

[5] Planck, M. (1913), *Wärmestrahlung (The Theory of Heat Radiation)*, English transl.: Masius, M. (1914; 1959) Dover Publ., NY.

[6] Shannon, C.E. (1948) A mathematical theory of communication, *Bell System Technical Journal*, 27: 379-423; 623-659.

[7] Brillouin, L. (1951b) Physical entropy and information II, *J. Appl. Phys.* 22(3) 338-343.

[8] Jaynes, E.T. (1957), Information theory and statistical mechanics, *Physical Review*, 106: 620-630.

[9] Tribus, M. (1961a), Information theory as the basis for thermostatics and thermodynamics, *J. Appl. Mech., Trans. ASME*, 28, 1-8.

[10] Tribus, M. (1961b), *Thermostatics and Thermodynamics*, D. Van Nostrand Co. Inc., Princeton, NJ.

[11] Kapur, J.N. & Kesevan, H.K. (1987), *The Generalized Maximum Entropy Principle (with Applications)*, Sandford Educational Press, Waterloo, Canada.

[12] Kapur, J.N. & Kesevan, H.K. (1992), *Entropy Optimization Principles with Applications*, Academic Press, Inc., Boston, MA.

[13] Tribus, M. (1969), *Rational Descriptions, Decisions and Designs*, Permagon Press, NY.

[14] Kapur, J.N. (1989b), *Maximum-Entropy Models in Science and Engineering*, John Wiley, NY.

[15] Kullback, S. & Leibler, R.A. (1951), On information and sufficiency, *Annals Math. Stat.* 22: 79-86.

[16] Kullback, S. (1959), *Information Theory and Statistics*, John Wiley, NY.

[17] Snickars, F. & Weibull, J.W. (1977) A minimum information principle: theory and practice, *Regional Science and Urban Economics* 7: 137-168.

[18] Jaynes, E.T. (1963), Information theory and statistical mechanics, *in* Ford, K.W. (ed), *Brandeis University Summer Institute, Lectures in Theoretical Physics, Vol. 3: Statistical Physics*, Benjamin-Cummings Publ. Co., 181; *also in* Jaynes, E.T., Rosenkratz, R.D. (ed.) (1983) *Papers on Probability, Statistics and Statistical Physics*, D. Reidel Publ. Co., Dordrecht, Holland, 39-76.

[19] von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics*, Princeton Univ. Press, Princeton.

[20] Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics, *Phil. Trans. Royal Soc. London* A 222, 309-368.

[21] Fisher, R.A. (1925) Theory of statistical estimation, *Proc. Camb. Phil. Soc.* 22: 700-725.

[22] Rényi, A. (1961) On measures of entropy and information, *Proc. 4th Berkeley Symp. Math. Stat. and Prob.* 1: 547-561.

[23] Kolmogorov, A.N. (1958) *Doklady Akademii nauk SSSR* 119: 861; (1959) 124: 754.

[24] Tsallis, C. (1988), Possible generalization of Boltzmann-Gibbs statistics, *J. Statistical Physics*, 52(1/2): 479-487.

[25] Tsallis, C. (2001), Nonextensive statistical mechanics and thermodynamics: historical background and present status, *in* Abe, S. & Okamato, Y. (eds), *Nonextensive Statistical Mechanics and its Applications*, Springer, Berlin, 3-98.

[26] Kaniadakis, G. (2001) Non-linear kinetics underlying generalized statistics, *Physica A* 296(3-

4): 405-425.

[27] Kaniadakis, G. (2002) Statistical mechanics in the context of special relativity, *Phys. Rev. E* 66(5): Art. No. 056125.

[28] Aczél, J. & Daróczy, Z. (1975) *On Measures of Information and their Characterization*, Academic Press, NY.

[29] Burbea, J. (1983), *J*-divergences and related concepts, *in* Kotz, S. & Johnson, N.L. (eds) *Encyclopedia of Statistical Sciences*, vol. 4., John Wiley, NY, 290-296.

[30] Papaioannou, T. (1985) Measures of information, *in* Kotz, S. & Johnson, N.L. (eds) *Encyclopedia of Statistical Sciences*, vol. 5., John Wiley, NY, 391-397.

[31] Kapur, J.N. (1983), A comparative assessment of various measures of entropy, *Journal of Information and Optimization Sciences* 4(3): 207-232.

[32] Kapur, J.N. (1984) A comparative assessment of various measures of directed divergence, *Advances in Management Studies* 3(1): 1-16.

[33] Behara, M. (1990), *Additive and Nonadditive Measures of Entropy*, John Wiley & Sons, NY.

[34] Arndt, C. (2001), *Information Measures: Information and its Description in Science and Engineering*, Springer Verlag, Berlin.

[35] Niven, R.K. (2005a), Exact Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac statistics, *Physics Letters A*, 342(4): 286-293.

[36] Niven, R.K. (2005b), Cost of *s*-fold decisions in exact Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac statistics, *http://arxiv.org/abs/cond-mat/0510128*.

[37] Tribus, M. & McIrvine, E.C. (1971), Energy and information, *Scientific American* 225: 179-188.

[38] Tolman, R.C. (1938) *The Principles of Statistical Mechanics*, Oxford Univ. Press, London.

[39] Schrödinger, E. (1952) *Statistical Thermodynamics*, Cambridge U.P., Cambridge.

[40] Hill, T.L. (1956) *Statistical Mechanics: Principles and Selected Applications*, McGraw-Hill, NY.

[41] Davidson, N. (1962) *Statistical Mechanics*, McGraw-Hill, NY.

[42] Eyring, H., Henderson, D., Stover, B.J. & Eyring, E.M. (1964) *Statistical Mechanics and Dynamics*, John Wiley & Sons, NY.

[43] Desloge, E.A. (1966), *Statistical Physics*, Holt, Rinehart & Winston, Inc., NY.

[44] Abe, R. (1975) *Statistical Mechanics*, University of Tokyo Press, Tokyo.

[45] McQuarrie, D.A. (1976) *Statistical Mechanics*, Harper & Row, NY.

[46] Atkins, P.W. (1982), *Physical Chemistry*, 2nd ed., Oxford University Press, Oxford, chap. 20, appendix A1.

[47] Jaynes, E.T. (1968), Prior probabilities, *IEEE Trans. Systems Science and Cybernetics*, SSC-4, 227-241.

[48] Jaynes, E.T. (Bretthorst, G.L., ed.) (2003) *Probability Theory: The Logic of Science*, Cambridge U.P., Cambridge.

[49] Jefferson, T.R. & Scott, C.H. (1986), The entropy approach to material balancing in mineral processing, *International Journal of Mineral Processing* 18: 251-261.

[50] Kapur, J.N. (1989a), Monkeys and entropies, *Bull. Math. Assoc. India* 21: 39-54.

[51] Stirling, J. (1730), *Methodus Differentialis: Sive Tractatus de Summatione et Interpolatione Serierum Infinitarum*, Gul. Bowyer, London, Propositio XXVII, 135-139.

[52] de Moivre (c.1733), *Miscellanea Analytica de Seriebus et Quadraturis*, J. Tonson & J. Watts, Londini.

[53] Kesevan, H.K. & Kapur, J.N. (1989), The generalized maximum entropy principle, *IEEE Trans. on Systems, Man, and Cybernetics* 19(5): 1042-1052.

[54] Kapur, J.N., Baciu, G. & Kesavan, H.K. (1995) The MinMax information measure, *Int. J. Systems Science* 26(1): 1-12.

[55] Yuan, L. & Kesavan, H.K. (1998) Minimum entropy and information measure, *IEEE Trans. Systems Man & Cybernetics C* 28(3): 488-491.

[56] Srikanth, M., Kesavan, H.K. & Roe, P.H. (2000) Probability density function estimation using the MinMax measure, *IEEE Trans. Systems Man & Cybernetics C* 30(1): 77-83.

[57] Topsøe, F. (2004) Entropy and equilibrium via games of complexity, *Physica A* 340: 11-31.

[58] Topsøe, F. (2005) Factorization and escorting in the game-theoretical approach to non-extensive entropy measures, *cond-mat/0510105*.

[59] Frieden, B.R. (2004), *Science from Fisher Information, An Introduction*, 2nd ed., Cambridge U.P., Cambridge.

[60] Szilard, L. (1929), Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen (On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings), *Zeitschrift für Physik* 53: 840-856; English transl: Rapoport, A. & Knoller, M. (1964), *in* Leff, H.S. & Rex, A.F. (1990), *Maxwell's Demon:*

*Entropy, Information, Computing*, Princeton Univ. Press, Princeton, NJ, 124-133.

[61] Wiener, N. (1948) *Cybernetics: or Control and Communication in the Animal and the Machine*, John Wiley, NY.

[62] Brillouin, L. (1951a), Maxwell's demon cannot operate: information and entropy I, *J. Appl. Phys.* 22(3) 334-337.

[63] Brillouin, L. (1953), The negentropy principle of information, *J. Appl. Phys.* 24(9): 1152-1163.

[64] Brillouin, L., (1949) Life, thermodynamics and cybernetics, *Am. Scientist* 37 554-568.

[65] Brillouin, L., (1950) Thermodynamics and information theory, *Am. Scientist* 38 594-599.

[66] Hartley, R.V.L. (1928), Transmission of information, *Bell System Technical Journal* 7(3): 535-563.

[67] Young, J.F. (1971), *Information Theory*, Butterworth, London, UK.

[68] Yaglom, A.M. & Yaglom, I.M. (1983), *Probability and Information*, D. Reidel Publishing Co., Dordrecht, Netherlands.

[69] Cover, T.M. & Thomas, J.A. (1991) *Elements of Information Theory*, John Wiley, NY.

[70] Niven, R.K. (2004), The constrained entropy and cross-entropy functions, *Physica A*, 334(3-4): 444-458.

[71] Shore, J.E. & Johnson, R.W. (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, *IEEE Trans. Information Theory* IT-26(1): 26-37.

[72] Tribus, M. (1988) An engineer looks at Bayes, *in* Erickson, G.J. & Smith, C.R. (1988) *Maximum-Entropy and Bayesian Methods in Science and Engineering*, vol. 1, Kluwer Academic Publ., Dordrecht, 31-52.

[73] Cox, R.T. (1961), *The Algebra of Probable Inference*, John Hopkins Press, Baltimore.

[74] Levine, R.D. (1980), An information theoretical approach to inversion problems, *J. Phys. A* 13: 91-108.

[75] Fowler, R.H. (1936), *Statistical Mechanics*, 2nd ed., Cambridge U.P., Cambridge.

[76] de Moivre, A. (1712), De mensura sortis, *Philosophical Transactions*, 27: 213-264, Prob. 8, English transl.: McClintock, B. (1984), *International Statistical Review*, 52(3): 229-262.

[77] Feller, W. (1957), *An Introduction to Probability Theory and its Applications*, 2nd ed., John Wiley, NY.

[78] Ratnaparkhi, M.V. (1985), Multinomial distributions, *in* Kotz, S. & Johnson, N.L. (eds) *Encyclopedia of Statistical Sciences*, vol. 5., John Wiley, NY, 659-665.

[79] Buteo, I. (1559), *Logistica*, Lugduni, 312-329; as quoted by Edwards [119].

[80] Bhaskara (c.1150), *Lilavati*, chap IV (section VI) and chap. XIII; English transl.: Colebrooke, H.T. (1817), John Murray, London.

[81] "IMDMI" (1636), *in* Mersenne, M. (1636), *Harmonicorum libri XII, Lutetiae Parisiorum*, Book VII, Prop. V, pp 118-119; see discussion by Edwards [119].

[82] Massieu, M. (1869) *Comptes Rendus* T. lxix.

[83] Gibbs, J.W. (1902), *Elementary Principles of Statistical Mechanics*, Dover Publ., NY.

[84] Einstein, A. (1902), Kinetische Theorie des Wärmegleichgewichtes und des zweiten Hauptsatzes der Thermodynamik (Kinetic theory of thermal equilibrium and of the second law of thermodynamics), *Annalen der Physik* 9: 417-433; English transl.: Beck, A. & Havas, P. (1989) *The Collected Papers of Albert Einstein, Vol. 2, The Swiss Years: Writings, 1900-1909*, Princeton Uinv. Press, NJ, 30-47.

[85] Einstein, A. (1903), Eine Theorie der Grundlagen der Thermodynamik (A theory of the foundations of thermodynamics), *Annalen der Physik* 11: 170-187; English transl.: Beck, A. & Havas, P. (1989) *The Collected Papers of Albert Einstein, Vol. 2, The Swiss Years: Writings, 1900-1909*, Princeton Uinv. Press, NJ, 48-67.

[86] Gibbs, J.W. (1875-1878), On the equilibrium of heterogeneous substances, *Trans. Connecticut Acad.* Oct. 1875-May 1876: 108-248; May 1877-July 1878: 343-524; *Am. J. Sci.* (1878) 16: 441-458; *also in* Gibbs, J.W. (1961) *The Scientific Papers of J. Willard Gibbs*, Dover Publ., Inc., NY, 55-371.

[87] Zwillinger, D. (2003), *CRC Standard Mathematical Tables and Formulae*, Chapman & Hall / CRC Press, Boca Raton, FL.

[88] Onsager, L. (1931a) Reciprocal relations in irreversible processes I, *Physical Review* 37: 405-462.

[89] Onsager, L. (1931b) Reciprocal relations in irreversible processes II, *Physical Review* 38: 2265-2279.

[90] Einstein, A. (1904), Zur allgemeinen molekularen Theorie der Wärme (On the general molecular theory of heat), *Annalen der Physik* 14: 354-362; English transl.: Beck, A. & Havas, P. (1989) *The Collected Papers of Albert Einstein, Vol. 2, The Swiss Years: Writings, 1900-*

*1909*, Princeton Uinv. Press, NJ, 68-77.

[91] Plastino, A. & Plastino, A.R. (1997), On the universality of thermodynamics' Legendre transform structure, *Phys. Lett. A*, 226, 257-263.

[92] Evans, R.B. (1969) *A Proof that Essergy is the Only Consistent Measure of Potential Work (for Chemical Substances)*, PhD thesis, Dartmouth College, NH (*unpub.*).

[93] Keenan, J.H. (1941) *Thermodynamics*, Wiley, NY.

[94] Keenan, J.H. (1951), Availability and irreversibility in thermodynamics, *Brit. J. Appl. Phys.* 2: 183-193.

[95] Rant, Z. (1956), Exergie, ein neues Wort fur, technische Arbeitsfahigkeit, *Forschung im Ingenieurwesen* 22(1): 36-37.

[96] Gaggioli, R.A. (1962) The concepts of thermodynamic friction, thermal available energy, chemical available energy and thermal energy, *Chem. Eng. Sci.* 17: 523-530.

[97] Ahern, J.E. (1980) *The Exergy Method of Energy Systems Analysis*, John Wiley, NY.

[98] Ayres, R.U., Ayres, L.W. & Martinás, K. (1998) Exergy, waste accounting, and life-cycle analysis, *Energy* 23(5): 355-363.

[99] Sciubba, E. & Ulgiati, S. (2005) Emergy and exergy analyses: Complementary methods or irreducible ideological options?, *Energy* 30: 1953-1988.

[100] Martinás, K. (1998) Thermodynamics and sustainability: a new approach by extropy *Periodica Polytechnica Ser. Chem. Eng.* 42(1): 69-83.

[101] Martinás, K. & Frankowicz, M. (2000) Extropy - reformulation of the entropy principle *Periodica Polytechnica Ser. Chem. Eng.* 44(1): 29-38.

[102] Gaveau, B., Martinás, K., Moreau, M. & Tóth, J. (2002) Entropy, extropy and information potential in stochastic systems far from equilibrium, *Physica A* 305: 445-466.

[103] Jaynes, E.T. (1978), Where do we stand on maximum entropy?, *in* Levine, R.D. & Tribus, M. (eds) (1978) *The Maximum Entropy Formalism*, MIT Press, Cambridge, MA, 15; *also in* Jaynes, E.T., Rosenkratz, R.D. (ed.) (1983) *Papers on Probability, Statistics and Statistical Physics*, D. Reidel Publ. Co., Dordrecht, Holland, 210-314.

[104] Jaynes, E.T. (1979), Concentration of distributions at entropy maxima, *in* Jaynes, E.T., Rosenkratz, R.D. (ed.) (1983) *Papers on Probability, Statistics and Statistical Physics*, D. Reidel Publ. Co., Dordrecht, Holland, 315-336.

[105] Boltzmann, L. (1896), Entgegnung auf die wärmetheoretischen Betrachtungen des Hrn. E.

Zermelo (Reply to Zermelo's Remarks on the Theory of Heat), *Annalen der Physik* 57: 773-384, Engl. transl., Brush, S.G. (1966), *Kinetic Theory, Vol. 2: Irreversible Processes*, Permagon Press, Oxford, 218-228.

[106] Einstein, A. (1905), Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt (On a heuristic point of view concerning the production and transformation of light), *Annalen der Physik* 17: 132-148; English transl.: Beck, A. & Havas, P. (1989) *The Collected Papers of Albert Einstein, Vol. 2, The Swiss Years: Writings, 1900-1909*, Princeton Uinv. Press, NJ, 86-103.

[107] Pearson, K. (1900), On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philos. Mag., 5th Series*, 50: 157-175.

[108] Fisher, R.A. (1924) The conditions under which $\chi^2$ measures the discrepancy between observation and hypothesis, *J. Royal Stat. Soc.* 87: 442-450.

[109] Kapur, J.N. & Saxena, H.C. (1969) *Mathematical Statistics*, 5th ed., S. Chand & Co., Delhi.

[110] Suyari, H. (2004a), Mathematical structure derived from the $q$-multinomial coefficient in Tsallis statistics, *cond-mat/0401546*.

[111] Suyari, H. (2004b), $q$-Stirling's formula in Tsallis statistics, *cond-mat/0401541*.

[112] Suyari, H., Tsukada, M. & Uesaka, Y. (2005), Mathematical structures derived from the $q$-product uniquely determined by Tsallis entropy, 2005 IEEE International Symposium on Information Theory, Adelaide, Australia, 4-9 September 2005.

[113] Suyari, H. (in press), Mathematical structures derived from the $q$-multinomial coefficient in Tsallis statistics, *Physica A*, accepted for publication.

[114] Jaynes, E.T. (1965) Gibbs vs Boltzmann entropies, *Am. J. Phys.* 33: 391-398.

[115] Jaynes, E.T. (1988), How does the brain do plausible reasoning?, *in* Erickson, G. J. & Smith, C. R. (eds.) *Maximum-Entropy and Bayesian Methods in Science and Engineering*, vol. 1, Kluwer, Dordrecht, 1.

[116] Bhandari, R. (1976) Entropy, information and Maxwell's demon after quantum mechanics, *Pramana* 6(3): 135-145.

[117] Jeffreys, H. (1932) On the theory of errors and least squares, *Proc. Royal Soc. London A* 138(834): 48-55.

[118] Jeffreys, H. (1961) *Theory of Probability*, 3rd ed., Clarendon Press, Oxford.

[119] Edwards, A.W.F. (2002), *Pascal's Arithmetical Triangle: The Story of a Mathematical Idea*, John Hopkins University Press, Baltimore.