Aspects of Information Theory

1. Chance and probability

How to define quantitatively "randomness" or "chaoticity" for a state and a process? Randomness has the connotation of "erratic" and "unpredictable." But the concept is difficult to quantify. For example, it is not obvious which of the two 100-particle configurations placed side by side in Fig. 1 is more random, more chaotic. In contrast to random states, the future behavior of a system in a stable, stationary state is predictable in its characteristic properties from the initial conditions and the appropriate equations of motion. But there is either not enough ***information*** available for a (classical) chaotic system, or what is available cannot be processed sufficiently well. ***Rather than with certainty, for chaotic and (fully or partially) random systems one has to work just with probabilities.*** In a different way, quantal systems are inherently always probabilistic, even for pure states.

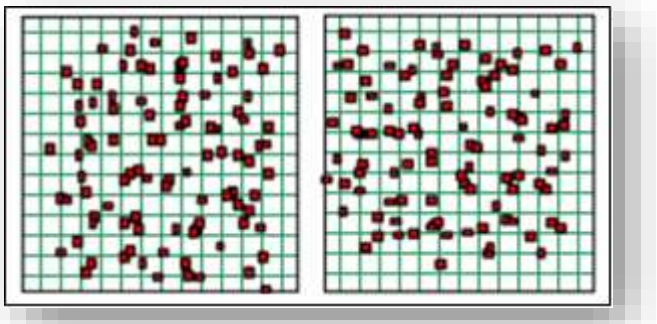


Figure 1: Two possible 2-dim systems of 100 particles each distributed differently over the available space.

In the absence of information, probability replaces certainty. Information theory provides probability as an objective link between randomness and certainty.

What is probability? One can think of repeated spinning of a roulette wheel, or the throwing of dice or coins, each ***a very large number of times N***, for example, $N = 1000$ throws of one dice or one throw each of $N=1000$ dice. These mechanical devices are capable of ***unpredictable, chaotic motion*** with seemingly random outcomes. Each set of *1000* independent observations (number hit on the roulette wheel, number on face of die,..) forms a "***statistical ensemble***" representing

possible configurations (states) of the system (wheel, die,..) populated in the process of spinning, throwing, etc.

The probability is defined in relation to a given set of experiments, a statistical ensemble representing a large number of possible system configurations. Then, in the above example the probability to throw a "*6*" with a perfect die is calculated from the number of times this face shows up in a throw. Say, one observes *N(6)=165* times a die with the face "*6*" on top. Then, the ***measured frequency*** of occurrence of this type of event is

$$P(6) = N(6)/N = 165/1000 \approx 1/6 \qquad (1)$$

In other trials of the same experiment, one may obtain different frequencies for the event "6", e.g., {*P(6)*}= {*173, 160, 155, 167, 182,…*}. This set of different numbers *P*(*6*) defines a ***distribution of frequencies***. The ***mean*** (= average = <*P*(*6*)=*N*(*6*)/*N*>) of all experiments is the experimental observation of interest for the mean frequency of face "*6*". The spread of this distribution is given by the ***variance***, which represents an experimental error or uncertainty.

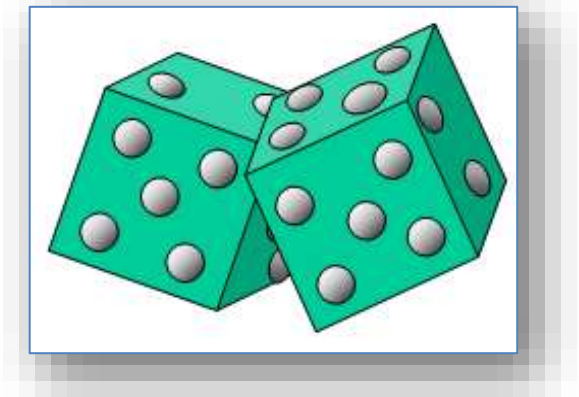


Figure 2: Study chance and probability in throwing of dice (P_i=*1/6*).

If the ***frequencies*** are measured for a large number of events (wheels spins, dice throws,..), i.e., for very large ensembles, the means of the frequency distributions approximate the probabilities to an arbitrary accuracy. Therefore, these mean frequencies are also called ***a posteriori probabilities***. For example,

$$P(6) = \lim_{N\to\infty} \frac{N(6)}{N} \overset{!}{=} \frac{1}{6} \qquad (2)$$

for the type of event considered in the example of dice throwing. In contrast, the ***a priori probability is defined*** as an idealization, a theoretical expectation: Assuming an ideal die with numbers *1, 2, 3, 4, 5, 6* printed on its otherwise identical faces, each face has the same *a priory* probability to show in any throw. ***Every number has the same chance to be on top in any given throw.***

If after one particular throw giving face "*6*", the same perfect die is rolled again, or if another perfect die is used in that throw, the ***chance*** (=*a priory probability*) to get a *6* (or any other number between *1* and *6*) is still *1/6*. Therefore, the probability to get a face "*6"* in either the first or the second throw is the sum of both,

$$P_{1\vee 2}(6) = P_1(6) + P_2(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \qquad (3)$$

Uncorrelated or mutually exclusive events (E_1 and E_2) are also called ***disjoint events***. The outcome of one trial (Event E_1) has no effect on the result of the next trial (Event E_2). The corresponding probabilities are independent of one another and add (as in Equ.(3)). This is the *sum rule* for disjoint (independent) probabilities.

On the other hand, one may ask what the (*a priori*) ***joint (simultaneous) probability*** or chance is for the throwing of *2* faces "6" in two independent throws. Obviously, in total there are $6\cdot 6 = 36$ possible combinations of two die faces. Therefore, the joint probability for obtaining two faces "*6"* simultaneously is *1* in *36,*

$$P_{1\wedge 2}(6\,|\,6) = P_1(6)\cdot P_2(6) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36} \qquad (4)$$

The simultaneous probability of two uncorrelated events is the product of the individual probabilities.

One can summarize the properties of the (a priori) probabilities P of events*:*

- The probability for an event is $0 \le P \le 1$
- The probability for any of the possible outcomes to occur is $\sum_i P_i = 1$
- The probability of an impossible event is zero, $P=0$.
- If two events (E) *1* and *2* are independent (disjoint, mutually exclusive), the probabilities of the sum ("or") event is the sum of the probabilities, $P_{1\vee 2} = P_1 + P_2$.
- If two events *1* and *2* are independent (disjoint, mutually exclusive), the probability for the simultaneous event is the product of the probabilities, $P_{1\wedge 2} = P_1 \cdot P_2$.
- If two events are not mutually exclusive, $P_{1\vee 2} = P_1 + P_2 - P_{1\wedge 2}$.

In addition, one has to consider cases of ***conditional (or marginal) probabilities***,

$$P\{E_1 \mid E_2\} := Probability\{E_1\},\ given\ E_2 \tag{5}$$

Here, event E_2 constitutes a boundary condition for all events E_1 to be considered but is not necessarily element in a causal chain. For example, one could ask for the probability that the solute in a solution precipitate, given that the solution is supersaturated in that component.

The following rules are evident for conditional probabilities:

$$P(E_1 \wedge E_2) = P\{E_2 \mid E_1\} \cdot P(E_1) = P\{E_1 \mid E_2\} \cdot P(E_2) \tag{6}$$

If E_1 and E_2 are independent,

$$P\{E_2 \mid E_1\} = P(E_2) \quad \text{and} \quad P\{E_1 \mid E_2\} = P(E_1) \tag{7}$$

What the effect of E_2 is on the conditional probability $P\{E_1 \mid E_2\}$ depends on the relation between the two classes of events, more accurately on the dependence of the probability domains ($P\neq 0$).

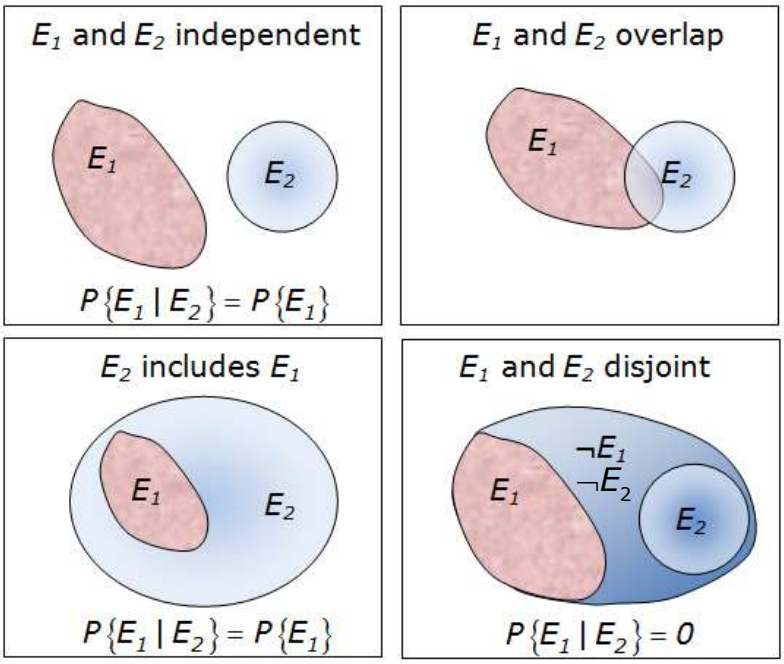


Figure 3: Probability domains for two different event classes that are (clockwise) independent and disjoint, overlapping, inclusive, disjoint and mutually exclusive.

In Fig. 3, several possible relations are illustrated between the domains of $P\{E_1\}$, $P\{E_2\}$, and $P\{\neg E_1\}$. The probability domain $P\{\neg E_1\}$ corresponds to the exclusion (non-occurrence) of E_1. An example of the conditional probability involving disjoint events is the probability that a dice throw has resulted in the face *"5"*, given that the result was an even number, i.e., in the present notation $P\{E_1 = "5" \mid E_2 = even\} = 0$.

If events E_1 and E_2 are independent of each other, or if the class of events E_2 includes that of E_1, then E_2 has no influence on the probability for E_1 and therefore $P\{E_1 \mid E_2\} = P\{E_1\}$ (Fig. 3). If the two events are disjoint, mutually exclusive, then $P\{E_1 \mid E_2\} = 0$.

The above equations (6) and (7) are equivalent to **Bayes' Theorem**,

$$P\{E_1 \mid E_2\} = P\{E_2 \mid E_1\} \cdot \frac{P(E_1)}{P(E_2)} \qquad (8)$$

2. Information and probability

The events and probabilities of interest in the present context refer to the ***occupation of states of physical systems***, the ***density of states occupied in the corresponding "phase space***." For example, one is interested in the intensity spectrum of a laser pulse, the size and growth rate of a population subgroup under certain conditions, the spread in the number of occupied cells in a cellular automaton, etc. The necessity for involving probabilities arises from the accessibility to the system of interest of different equivalent states and the lack of knowledge, in which of these states the system resides at any given time.

The difference between the probability to find a system in a certain configuration and certainty is caused by the lack of prior information, which can be large if the accessible phase space is large and the motion is fairly unrestrained ("disordered" or "chaotic"). In order to be able to make predictions for the evolution of the system, it is important to quantify this missing information based on *a priori* properties of the system states and for a range from orderly to random modes. The following discussion illustrates the intrinsic connections between probability and information for systems that have countable discrete states. While this latter feature is observed for all bound states of microscopic systems (e.g., molecules, atoms, nuclei, nucleons,..), the concept can be extended to continuous states invoking minimum quantal phase space cells. The principles developed below have therefore a rather broad range of validity.

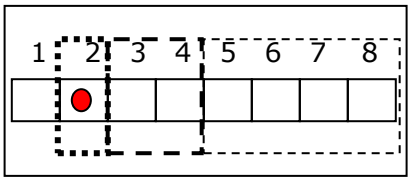


Figure 4: Automaton with one cell (#2) occupied.

The connection between probability and lack of information (Claude Shannon, 1948) can be illustrated with cellular automata. Here, each cell can be viewed to represent a cell in the phase space of a physical system. A cell could be occupied by a particle or could be unoccupied. Consider the one-dimensional automaton of *8* cells pictured in Fig. 4. If it were known that only

one specific cell were accessible, say cell *#1*, the sole particle of the system would be known to be in that cell with certainty (P(*2*)=*1*). ***No information would be missing***.

Now assume, all *N=8* cells were identical and equally accessible to the particle. Each cell has therefore two possible states (occupied=*1* or not occupied=*0*). In total, there are $N=2^3$ possible states of the automaton. Since the cells are identical, they have *equal a priory probabilities* to be occupied, and the actual location of the particle is unknown and "completely uncertain." The question is how much information is missing prior to an observation, or is obtained when the location of the particle is revealed. One only knows that the ***missing information is at a maximum <u>if</u> all states are a priori equally likely***. If not all cells were equally likely to be occupied, less information would be missing.

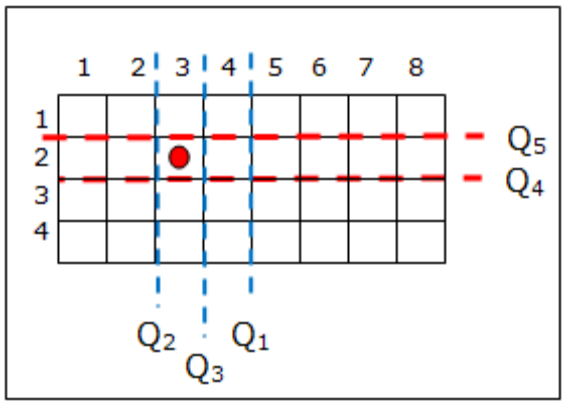
As a measure of the magnitude of the missing information, one can take the number of inquiries necessary to determine the state of the automaton, i.e., determine which cell is occupied. There are many more or less efficient strategies one could apply to obtain the missing information. The simplest consists in dividing the space into successively smaller halves. One could start with asking whether the particle is in the right half (cells *#5-8*). Since in the example of Fig. 4, is "no" that half is eliminated.

The subsequent second question (Q_2) then results in a further increased knowledge that the particle must be in the quarter of the space containing cells #1 and #2. Finally, Question Q_3 definitively locates the particle in cell #2. While it is possible that, by chance, one could have guessed that cell correctly, it is always possible to locate the particle correctly within an array of $N=2^3$ cells with

$$I(N=8) = Log_2 N = 3 \qquad (9)$$

binary (yes/no) inquiries. To see whether or not this is a systematic feature, consider a ***2*-dim automaton** of ***N = 32 = 2^5* identical**

Figure 5: 2D automaton grid with one cell out of *N=32* occupied.



cells accessible to the single particle that makes up the test system (see Fig. 5). Using the same strategy, first in the horizontal dimension and then in the vertical dimension, one subdivides the *4x8* cell array into successively smaller halves and queries in which of the halves the particle is located. Now it takes exactly *5* binary (yes/no) questions to eliminate all gaps in the knowledge of the state of the system. The *5* necessary questions are illustrated in Fig. 5 by labels and dividing lines. Again, one finds the relation between the measure (I) of missing information and the number of possibilities (N) given in analogy to Equ. (9) as

$$I(N) = Log_2 N \tag{10}$$

Again, the information is encoded in terms of the minimum required number of probings with binary "yes" or "no" answers. It is encoded in "bits."

Since in the above example there are N equally probable possibilities for the particle to reside, the probability for each cell to house the particle is the inverse,

$$p = \frac{1}{N} \tag{11}$$

The information in Equ. (10) can then be expressed also in terms of this probability,

$$I(N) = Log_2 N = -Log_2 p \tag{12}$$

Obviously, the missing information scales trivially with increasing number (N) of states accessible to the particle. Therefore, a natural measure of the relative magnitude of the missing information is the above quantity $I(N)$ per single-particle state (cell):

$$s = \frac{I(N)}{N} = -p \cdot Log_2\, p \tag{13}$$

In other words, the quantity *I(N)* is a numerical measure of maximally possible randomness of the single-particle configuration. Chaotic populations result, if the underlying process has completely unrestricted access to all system configurations, in which case the observed probabilities model the *a priori* ones.

The quantity *s*, which is also known as ***statistical entropy***, is the maximum of the information missing to identify the single-particle state that will be found occupied in an observation. In other words, this information is gained, once the location of the particle is revealed. Since the expression shown in Equ. (13) is normalized per state *N*, one only needs only the probabilities *p* for its evaluation but not the absolute number of states.

The above considerations, which are based on the binary (yes/no) nature of the property "occupied=bit 1" vs. "not occupied=bit 0", led to expressions of Equs. (10) and (13) in terms of the logarithm to basis *2* ("*bits*"). Instead of this function, one often uses the natural logarithm $\ell n\, x = Log_e x$ ("*nats*"). The corresponding equations

$$I(N) = k\ \ell n\, N \tag{14}$$

and

$$\frac{I(N)}{N} = -k\, p \cdot \ell n\, p \tag{15}$$

then differ from the original ones by a conversion factor *k* for the logarithms. The information is expressed in different "units" (***nats*** vs. ***bits***). However, it is important that these units really have no dimension such as length, time, energy, etc. *Values given for information or statistical entropy simply reflect the number of binary questions required to establish the state of the system of interest with certainty.*

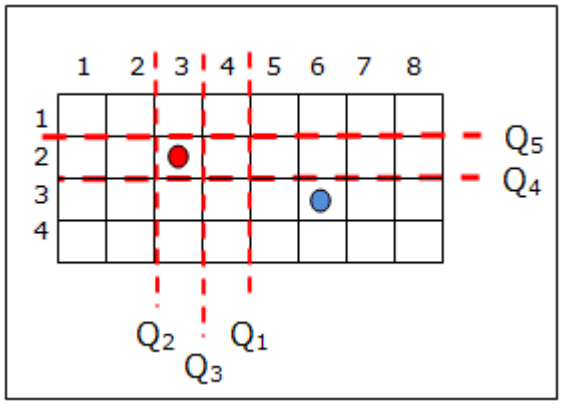


Figure 6: Two-dimensional two-particle system.

The statistical systems of physical interest have typically many degrees of freedom, for example a large number ($n \sim 10^5$ - 10^{23}) of particles that can be distributed over N single-particle states. This situation is illustrated in Fig. *6* for just two independent particles (*1, 2*), thought to be distributed randomly (without preference) over all possible states. The states (cells) have ***equal a priori probabilities*** and are therefore equally probable to receive either of the two particles. The missing information on either particle is evaluated as explained above for the one-particle system. It is obvious that, for N ***single-particle states***, the total ***number of two-particle states*** is given by $\Omega = N_1 \cdot N_2 = N^2$. Therefore, the number of binary queries necessary to identify a 2-particle configuration equals twice the number needed to pin down one particle:

$$I(\Omega) = I(N_1 \cdot N_2) = k\left[\ell n N_1 + \ell n N_2\right] \qquad (16)$$

In the following, this problem is generalized to an arbitrary number of individual particles and an arbitrary number of single-particle states. The conversion constant is set to $k=1$, implying the information has been scaled by a factor of $1/k$.

Using mathematical combinatorics, one can easily expand the above method of information gathering to an arbitrary large number M of particles or other objects, and/or to many degrees of freedom. All that is required is that the degrees of freedom are independent and not coupled to each other and that the states are discrete and countable. Then, the total number Ω of states is given by the product of the numbers Ω_i of states for independent degrees of freedom $i=1,...,M$,

$$\Omega = \prod_{i=1}^{M} \Omega_i \qquad (17)$$

For the explicit example below, $M=2$ is used for convenience.

For specificity, let N_1 be the number of independent individual particles that can occupy ***with equal a priory probability*** N_1 out of the $N > N_1$ available single-particle states. The quantity Ω_1 equals the number of N_1-particle states that can be formed out of the N available single-particle states. The problem to be solved is to calculate the number (Ω_1) of ways the N_1 particles can occupy N_1 different but equivalent single-particle states, while $N_2=N-N_1$ states remain unoccupied. For classical objects, this number is given by the binomial coefficient

$$\Omega_1 = \binom{N}{N_1} = \frac{N!}{N_1!\cdot(N-N_1)!} = \frac{N!}{N_1!\cdot N_2!} \qquad (18)$$

Here, $N_2 = (N - N_1)$, and $N_1 + N_2 = N$ is the normalization. Because of the symmetry of expression(18), the quantity Ω is also equal to the number of possibilities to select N_2 objects (unoccupied states) out of N total.

In general, any group of N classical objects can be rearranged by permutation in $N! = 1\cdot 2\cdot 3\cdots N$ different ways, i.e., it has $N!$ permutations. Then, the number of possible "partitions" of N of the type $N_1, N_2, \ldots, N_M$ with each N_i differing only by permutation of its objects is given by $N_{1,..,M} = \prod_{i=1}^{M} N_i!$. For $\sum_i N_i = N$, the number Ω of possible configurations of $\{N_1, N_2, \ldots, N_M\}$, i.e., the number of possible ***partitions of the number N***, is equal to the number of total permutations $N!$ divided by $N_{1,\ldots,M}$,

$$\Omega(N;\ N_1,\ldots,N_M) = \frac{N!}{N_1!\cdots N_M!} \qquad (19)$$

This realization provides a straightforward way to extend Equ. (18) to an arbitrary number of subgroups of states, distinguished, e.g., by the object located on it or any other property such as energy, spin, etc.

The problem is particularly simple for macroscopic thermo-dynamical systems, where the number of particles is large ($N_1 \gg 1$) and the

number of available states is even larger (often $N >> N_1$). Then, one can always neglect numbers of the order of *1* in comparison with either N_1 or N_2. In order to evaluate Equ. (18), one may then use Stirling's formula,

$$\ell n(n!) \approx n \cdot \left[\ell n(n) - 1\right] \tag{20}$$

As before, one can define as a measure of information the number of ***bits*** or ***nats*** making up the number of *equal a priori* possibilities. From Equs. (14) and (18), one obtains in the limit of large numbers (» *1*),

$$\begin{aligned} I := \ell n\Omega = \quad & \ell n N! \quad - \quad \ell n N_1! \quad - \quad \ell n N_2! = \\ & \approx \overbrace{N(\ell n N - 1)} - \overbrace{N_1 \ell n(N_1 - 1)} - \overbrace{N_2 \ell n(N_2 - 1)} \\ & = N\ell n N - N_1 \ell n N_1 - N_2 \ell n N_2 \\ & = (N_1 + N_2)\ell n N - N_1 \ell n N_1 - N_2 \ell n N_2 \end{aligned} \tag{21}$$

or, since $N_1 + N_2 = N$,

$$I \approx -N_1 \ln \frac{N_1}{N} - N_2 \ln \frac{N_2}{N} \tag{22}$$

This yields an information per total number *N* of states,

$$S = \frac{I}{N} \approx -\frac{N_1}{N} \ln \frac{N_1}{N} - \frac{N_2}{N} \ln \frac{N_2}{N} = -\sum_{n=1}^{2} p_n \ln p_n \tag{23}$$

Here, the quantities

$$p_n = \frac{N_n}{N} < 1; \qquad (n = 1,2) \tag{24}$$

are the relative probabilities to find in a measurement any N_1 of the *N* possible single-particle states occupied ($n=1$), or any of the N_2 single-particle states not occupied ($n=2$). The total number *N* of single-particle

states has been partitioned into two groups of N_1 and $N_2=N-N_1$, each found realized with the respective fractional weights.

If there are more than 2 subgroups of states, e.g. distinguished by the objects they carry, one can generalize Equ. (23) to an arbitrary number of M subgroups (or ***partitions of the number N***) of states (or objects),

$$S = \frac{I}{N} = -\sum_{n=1}^{M} p_n \ln p_n; \;\; with \quad \sum_{n=1}^{M} p_n = 1 \tag{25}$$

The equation on the right in (25) is the trivial normalization of constant total probability. This formalism accommodates any partition of the total number of individual states into subgroups. For example, it provides the same formulas when the index n is redefined to number ($n=n_1 \le M=N$) the individual single-particle states, or the 2-particle states ($n = n_{12} \le M = \binom{N}{2}$), etc. In other words, expression (25) also describes the information/entropy for M different subgroups of occupied states out of the Ω available to the M particles.

Of specific interest is the trivial case, in which one particle can access all possible $\Omega = N$ single-particle states with equal *a priori* probability. Then, Equ. (21) yields

$$I = \ell n\, \Omega = -\ell n\left(1/\Omega\right) \tag{26}$$

or, with a constant, equal (*a priori*) probability per state,

$$p_n = 1/\Omega; \quad n = 1,\ldots,\Omega \tag{27}$$

$$s = \frac{I}{\Omega} = -\frac{1}{\Omega}\ell n \frac{1}{\Omega} = -p_n \ell n\, p_n \tag{28}$$

The total information, summed over all states, becomes

$$S = I = -\sum_{n=1}^{\Omega} p_n \ell n\, p_n \qquad (29)$$

3. Information and partition of probability

The formal information I (Equ. (25)) gained through the acquisition of knowledge about which of any of the possible configurations is realized, represents the information that the totality of *nats* or *bits* of the Ω states can carry. This information is numerically equal to the number (= $\ell og_2\,\Omega$) of questions that have to be posed in order to obtain certainty.

It is plausible that any asymmetry, any restriction in the probabilities for the various system configurations would imply a bias, which reduces the information content. For example, a 12-bit computer cell with two broken bits can not carry more information than a 10-bit word. Therefore, an ***equipartition of the total probability***

$$P = \sum_i p_i, \qquad p_i = p = const. \qquad (30)$$

among all system states (configurations) maximizes the information and the statistical entropy (cf. Equ. (13)). In such a situation, where all configurations have an equal *a priori* probability, a many times repeated experiment is expected to exhibit every configuration with the same equal *a posteriori* (empirical) probability. For example, measuring 10000 times the same system with 3 equally probable configurations will show each one of them approximately 3300 times (1/3 of the total number of interrogations).

Mathematically the partition of the total probability among the system configurations that corresponds to maximum information (entropy) can be obtained by varying the probabilities p_n under the constraint of the normalization of all probabilities. This task is achieved efficiently with the ***method of Lagrange multipliers***. To illustrate the method, consider a 1-dimensional function *f(x).* Here a "constrained maximum"

of *f(x)*, under the constraint *g(x)=c=const.*, is found by searching for the maximum of the related function

$$\tilde{f}(x) = f(x) - \lambda\left[g(x) - c\right] \tag{31}$$

where λ is an arbitrary constant. Obviously, only along a path where the constraint *g(x)=c* is fulfilled, are the functions $\tilde{f}$ and *f* identical, $\tilde{f}(x) = f(x)$.

In the case of interest here, the constraint is given by the normalization condition of the total probability for any of the Ω configurations,

$$g(p_1,\ldots,p_\Omega) = \sum_{i=1}^{\Omega} p_i = 1 \tag{32}$$

Then, in the usual way, the maximum of the information is found by setting to zero all first derivatives of $\tilde{f}$ with respect to individual probabilities p_n (for configuration *n* to be occupied, $n=1,\ldots,\Omega,$),

$$\frac{\partial}{\partial p_n}\left\{-\sum_{m=1}^{\Omega} p_m \,\ell n\left(p_m\right) + \lambda\left(\sum_{i=1}^{\Omega} p_i - 1\right)\right\} = 0 \tag{33}$$

The normalization condition of Equ. (32) has been multiplied by a yet undetermined but constant ***Lagrange multiplier*** λ and added to the function to be maximized. This procedure yields

$$-\ell n\left(p_n\right) - p_n \cdot \frac{1}{p_n} + \lambda = 0 \qquad n = 1,\ldots,\Omega \tag{34}$$

or

$$\ell n\left(p_n\right) = \lambda - 1 \tag{35}$$

and

$$p_n = e^{(\lambda-1)} = p = const. \tag{36}$$

implying ***an equal (a priori) probability for all n***. From the normalization condition $\sum_{n=1,..,\Omega} p_n = 1$ it follows immediately that

$$p_n = p = 1/\Omega = const \tag{37}$$

and

$$1 = \sum_{n=1}^{\Omega} p_n = \sum_{n=1}^{\Omega} e^{(\lambda-1)} = \Omega \cdot e^{(\lambda-1)} \tag{38}$$

which determines the Lagrange multiplier. This is the condition for ***maximum missing information*** concerning the locations of the particles of interest. It also signifies the situation for a ***maximum of information gained***, when the occupation of the single-particle states is disclosed.

In analogy to Equ. (25), the ***statistical entropy*** used in statistical mechanics for a system of N particles populating a set of Ω configurations ("micro-states") is defined as

$$S = -k_B \sum_{n=1}^{\Omega} p_n \ell n(p_n) \geq 0 \tag{39}$$

where $k_B = 1.38 \cdot 10^{-23}\ J/K$ is the universal ***Boltzmann constant***, which endows this information quantity with a non-trivial but unnecessary dimension that can obscure the real meaning of this important observable. If the system is interrogated N times and configuration n is observed N_n times out of N, the *a* ***posteriori (empirical) probabilities*** are determined by $p_n = N_n/N$. These probabilities fulfill the normalization condition

$$\sum_{n=1}^{\Omega} p_n = \sum_{n=1}^{\Omega} \frac{N_n}{N} = 1 \tag{40}$$

Following the same arguments as above for the information, one finds that the ***entropy is maximized for an equal population of all***

configurations (micro-states) with equal a priori probabilities $p_n = 1/\Omega$. In this case of independent individual objects that occupy ***with equal a priory probability*** all Ω available states, Equ. (14) translates into the ***maximum statistical entropy***,

$$S = S_{max} = k_B \cdot \ell n\,\Omega = -k_B \sum_{n=1}^{\Omega} p_n\, \ell n\left(p_n\right) \geq 0 \qquad (41)$$

The left hand side of this equation is the famous *Boltzmann Equation* relating the number (Ω) of available micro-states to the phenomenological entropy, a state function defining macroscopic system states in phenomenological thermodynamics. The right hand side closes contact to the microscopic information content. The constant k_B, providing the information observable with an artificial dimension, makes more sense in phenomenological thermodynamics. The entropy is naturally bounded by the two fixed limits,

$$0 \leq S \leq S_{\max} \qquad (42)$$

where S_{max} is attained for equal a priori probabilities, for any physical system it is itself a distribution that can be characterized by average expectation value, fluctuations and higher moments.

In a similar fashion, the method utilized above for evaluating the maximum information/entropy under the constraint of an equipartition of the total *a priori* probability can be extended to other constraints. For example, for a multi-particle system it is important how the system energy E is distributed among all Ω configurations, i.e., micro-states. For an isolated ("**micro-canonical**") system the total energy E is conserved, as are other observables. Then, all members of the set of possible equivalent configurations must have exactly the same energy,

$$E_i = E \qquad i = 1,\ldots,\Omega \qquad (43)$$

Otherwise, the configurations would not be completely equivalent, i.e., not have the same *a priori probabilities*. In a statistical ensemble, this

fixed energy *E* would be distributed over all members of the ensemble with *equal a priori probabilities*. Systems that are not completely isolated and allow some exchange of energy with their surroundings are called "**canonical**." Here, the energies of the various configurations are subject to ***fluctuations about the averages***, <E_i>. ***Exactly equal a priori probabilities cannot be expected for these configurations.*** At most some energy averages can be well defined and assumed to have approximately the same value for every configuration:

$$\langle E_i \rangle = \sum_{i=1}^{\Omega} E_i \cdot p_i =: E \tag{44}$$

Here, $\langle E_i \rangle$ is the weighted average (mean) taken over all configurations. The condition that only small variations should exist between the mean energies E_i of the configurations implies that the ***mean square deviation (variance)*** of these energies

$$\sigma_E^2 = \sum_{i=1}^{\Omega} \left(E_i - \langle E_i \rangle\right)^2 \cdot p_i > 0 \tag{45}$$

be small compared to the average, i.e., $\sigma_E \ll E$. For an isolated, micro-canonical system, there are no energy fluctuations, $\sigma_E = 0$.

Considerations of the ***maximum constrained information*** will reveal whether such a situation is possible and what the *a priori probabilities* p_i would look like. Certainly, ***because of an additional constraint, for a given energy E, the information/entropy cannot exceed the one obtained for a micro-canonical system*** with a minimum of constraints (total probability normalized).

A search for the maximum information/entropy has to take into account now two boundary conditions, Equ. (32) and Equ. (44). Therefore, the maximization condition (33) has to be extended to

$$\frac{\partial}{\partial p_n}\left\{-k_B \sum_{m=1}^{\Omega} p_m \, \ell n(p_m) + \lambda_1 \left(\sum_{i=1}^{\Omega} p_i - 1\right) + \lambda_2 \left(\sum_{j=1}^{\Omega} E_j p_j - E\right)\right\} = 0 \tag{46}$$

with two Lagrange multipliers, λ_1 and λ_2. For the constraint maximum information/entropy one now obtains the condition

$$-k_B\left(\ell n(p_n)+1\right)+\lambda_1+\lambda_2 E_n = 0 \qquad (47)$$

This result implies that

$$p_n = exp\left\{\frac{\lambda_1+\lambda_2 E_n}{k_B}-1\right\} \qquad n=1,\ldots,\Omega \qquad (48)$$

where λ_1 / k_B is dimensionless and λ_2 / k_B is an inverse energy. Obviously, the probabilities are not equal but depend on the energies E_n of the corresponding states. The normalization condition is now written as

$$1 = \sum_{n=1}^{\Omega} p_n = e^{\left(\frac{\lambda_1}{k_B}-1\right)} \cdot \sum_{n=1}^{\Omega} e^{\frac{\lambda_2}{k_B}E_n} \qquad (49)$$

a product of a constant and an energy sum. Obviously, the constant factor in Equ. (49) is equal to the inverse of the sum over the individual energy terms,

$$Z = e^{\left(1-\frac{\lambda_1}{k_B}\right)} = \sum_{n=1}^{\Omega} e^{\frac{\lambda_2}{k_B}E_n} =: \sum_{n=1}^{\Omega} e^{-\beta\cdot E_n} \qquad (50)$$

with the definition $\beta = -\lambda_2/k_B > 0$, the inverse of a characteristic energy. This function ***Z=Z(β)*** is also known as the ***(canonical) partition sum***. According to Equ. (50) it can be cast both into a closed form (left) and as a sum over all configurations. Hence the normalization condition is recast as

$$\sum_{n=1}^{\Omega} p_n = \frac{1}{Z}\cdot\sum_{n=1}^{\Omega} e^{-\beta\cdot E_n} = 1 \qquad (51)$$

This is a general result, valid for any number of configurations, their energy spectra (E_n) and varied system parameters λ_1 and λ_2.

Often groups of several (ϖ) states have the same energies, i.e., they are ***energy degenerate*** and bunched at some energy levels *E, E', etc*. If the degeneracy (number of states at energy level *E*) is given by the function $\varpi(E)$, the partition function in Equ. (50) can be written

$$Z(\beta) = \underbrace{e^{-\beta\cdot E} + \ldots + e^{-\beta\cdot E}}_{\varpi(E)\ times} + \underbrace{e^{-\beta\cdot E'} + \ldots + e^{-\beta\cdot E'}}_{\varpi(E')\ times} + \underbrace{\ \ldots\ }_{\varpi(\ldots)} = \sum_{E} \varpi(E)\cdot e^{-\beta\cdot E} \qquad (52)$$

According to Equ. (48), there is a term-by-term equivalence in Equ. (51). One therefore concludes that the ***a priori probabilities for canonical system configurations are not equal but dependent on the energy.*** Configuration by configuration, one has the normalized probability

$$p_n = \frac{1}{Z}\cdot e^{-\beta\cdot E_n} \qquad (53)$$

Accordingly, the probabilities for the populations $p(E)$ of energy levels *E* are given by

$$p(E) = \frac{\varpi(E)}{Z}\cdot e^{-\beta\cdot E} \qquad (54)$$

The requirements that the probabilities must be normalizable and that variations between the mean energies of equivalent (similar probabilities) be small suggests that the inverse-energy parameter β be positive, $\beta > 0$. Then, the populations for system configurations decrease exponentially with their energy, Configurations with extreme energies are simply not significantly populated. In fact, for thermodynamic systems independent considerations discussed further below show a relation of the parameter with the "canonical temperature" *T,* i.e., $\beta = 1/(k_B T)$. This implies that the information contained in such a system is incomplete at any temperature, reduced due to the decreased probability for energetic states.

The partition sum contains all relevant physical information on the system. *Z* ***is a generating function for the system probability distribution.*** This feature can be demonstrated by the following examples.

The derivative of $\ell n(Z)$ with respect to the energy E_i projects the probability of configuration *i* out of the partition sum,

$$-\frac{\partial}{\partial E_i}\ell n(Z) = \frac{-1}{Z}\frac{\partial}{\partial E_i}\sum_{n=1}^{\Omega} e^{-\beta\cdot E_n} = \frac{\beta}{Z}e^{-\beta\cdot E_i} = \beta p_i \tag{55}$$

Here the chain rule

$$\frac{\partial}{\partial x}\ell n(Z) = \frac{1}{Z}\frac{\partial}{\partial x}Z \tag{56}$$

has been used to generate the required normalization factor *1/Z*. Similarly, taking the derivative of $\ell n(Z)$ with respect to -β produces the mean energy per configuration:

$$-\frac{\partial}{\partial \beta}\ell n(Z) = \frac{-1}{Z}\frac{\partial}{\partial \beta}\sum_{n=1}^{\Omega} e^{-\beta\cdot E_n} = \frac{1}{Z}\sum_{n=1}^{\Omega} E_n e^{-\beta\cdot E_n} = \sum_{n=1}^{\Omega} E_n\cdot p_n = \langle E_n\rangle = E \tag{57}$$

Taking the result from Equ. (47), multiplying by p_n and summing over all configurations yields a connection between entropy, partition function and mean energy per configuration

$$\begin{aligned} 0 &= -k_B\sum_{n=1}^{\Omega} p_n\left\{\left(\ell n(p_n)+1\right)+\lambda_1/k_B+\left(\lambda_2/k_B\right)E_n\right\} = \\ &= S - k_B\sum_{n=1}^{\Omega} p_n + \lambda_1\sum_{n=1}^{\Omega} p_n + \lambda_2\sum_{n=1}^{\Omega} p_n E_n \\ &= S-\left(k_B-\lambda_1\right)-k_B\cdot\beta\cdot E = S-k_B\cdot\left[\ell n(Z)-\beta\cdot E\right] \end{aligned} \tag{58}$$

This results finally yields an expression for the macroscopic (mean) information/entropy in terms of the partition function *Z* and the expectation value of the energy,

$$S/k_B = \ell n\, Z + \beta\cdot E \tag{59}$$

Equivalently, one can write for the partition function for a canonical system,

$$Z = e^{S/k_B}\cdot e^{-\beta\cdot E} \tag{60}$$

This function replaces that for an isolated system, which according to Equ. (41), is simply the number of accessible states,

$$\Omega = e^{S/k_B} \tag{61}$$

with the dimension-less statistical entropy S/k_B counting the number of "nats" measuring the size of the state space.

So far, the meaning and value of the parameter β appearing in the canonical partition function, have remained hidden. However, for any system obeying the Equ. (59), the parameter obeys the relation

$$\beta = \frac{\partial\left(S/k_B\right)}{\partial E} \quad or \quad \frac{\partial E}{\partial S} = \frac{1}{k_B \cdot \beta} \tag{62}$$

It can therefore be evaluated for a system of interest, given specific relations between state energies and probabilities. Note that the derivatives in Equ. (62) are partial derivatives testing explicit dependencies, to be taken while keeping other coordinates constant.

4. Illustrations of partition functions

A detailed evaluation of the partition function is necessary for an interpretation of macroscopic observations in terms of the microscopic structure of a system, e.g., in terms of the internal energy spectrum. The task can be very demanding for quantal multi-particle systems with coupled degrees of freedom and correlated particles, e.g., for fermionic systems where the individual particles are indistinguishable and subject to the Pauli Exclusion Principle. Particle correlations are important for high particle densities in configuration or momentum space but loose efficiency at low densities and/or total internal energies. On the other hand, many classical d.o.f. such as molecular or nuclear rotations and vibrations are independent at low excitations but influence each other at higher energies. In the following, a few observations relevant to

decoupled d.o.f. are made in order to illustrate basic structure of partition functions for various d.o.f.

For a system with multiple independent degrees of freedom, for example, a molecule with a set (*i=translational, rotational, vibrational, electronic, nuclear,...*), the total number of states Ω for is the product of the corresponding numbers Ω_i for the individual d.o.f. Its total energy $E = \langle E \rangle$ is the sum over the individual energies $E_i = \langle E_n(i) \rangle$. Therefore, the total partition function Z is the product of the individual functions Z_i corresponding to each of the d.o.f.,

$$Z(\beta) = \prod_i Z_i(\beta) = \prod_i \sum_n e^{-\beta \cdot E_n(i)} \tag{63}$$

Here, the energies $E_n(i)$ run over the entire energy spectrum associated with the i th d.o.f. In other words, as long as correlations can be neglected (quasi-classical, Boltzmann approximation), the partition function for such a system can be written as

$$Z = Z^{trans} \cdot Z^{rot} \cdot Z^{vib} \cdot Z^{electr} \cdot Z^{nucl} \ldots \tag{64}$$

Furthermore, for a quasi-classical N-particle molecular system, neglecting correlations, each partial partition function is a product of identical single-particle partition functions z_i. For example, for translational motion in 3D space $\{x, y, z\}$, the s.p. partition function can be approximated by

$$z^{trans} = \sum_{n_x, n_y, n_z} e^{-\beta \cdot \left(\varepsilon_{n_x} + \varepsilon_{n_y} + \varepsilon_{n_z} \right)} = \left(\sum_i e^{-\beta \cdot \varepsilon_i} \right)^3 \tag{65}$$

Here, the s.p. energy scheme of a particle in an infinite cubic 3D box of side length $a\tau\equiv$ is adopted. Such an infinite box accommodates the unrestricted translational motion of a free particle of

mass *m*. For mathematical reasons, one adopts first a finite box but lets its dimension grow indefinitely in the final results.

For a finite side length the particle-in-a-box energy eigen values are given by a set of integer quantum numbers n_i

$$\varepsilon_i = \frac{h^2}{8ma^2} \cdot n_i^2 \tag{66}$$

Performing the transition to the infinite box, $a \delta \equiv$, the summation over discrete quantum numbers n_i in the partition sum can be replaced by an integral over continuous quantum numbers *n:*

$$z^{trans} = z^{trans}_{a\to\infty} = \lim_{a\to\infty}\left[\int_0^\infty dn e^{-\beta\cdot\frac{h^2}{8ma^2}\cdot n^2}\right]^3 = \left[\sqrt{\frac{2\pi}{\beta\cdot h^2}m}\cdot a\right]^3 \tag{67}$$

Since the volume is given by $V = a^3$, the translational partition function can also be written as

$$z^{trans} = \frac{V}{\lambda^3_{therm}} \propto \beta^{-3/2} \tag{68}$$

In this expression, λ_{therm} is the so-called thermal wave length,

$$\lambda_{therm} = \sqrt{\frac{\beta\cdot h^2}{2\pi m}} \tag{69}$$

With this detail knowledge of the s.p. translational partition function, one can now calculate the expectation (mean) value of a particle in a canonical system. From Equ. (57) one has

$$\langle\varepsilon\rangle = -\frac{\partial}{\partial\beta}\ell n\, z = -\frac{\partial}{\partial\beta}\ell n\,\beta^{-3/2} = \frac{3}{2}\cdot\frac{1}{\beta} \tag{70}$$

In Equ. (70) the fact has been used that there is only one factor in the function *z(β)* that actually depends on the parameter β. Obviously, the mean energy of a free particle in a canonical system can, and has been, measured to be

$$\langle \varepsilon \rangle = \frac{3}{2} \cdot k_B \cdot T \tag{71}$$

In the development of thermodynamics, the mean kinetic energy of a free particle has been identified (by convention) with the product of Boltzmann constant k_B and ***temperature T***. Therefore, one has to identify,

$$\beta = \frac{1}{k_B \cdot T} \tag{72}$$

The heretofore unknown model parameter β has now been linked to experimental observation. It is an inverse energy which at room temperature T=300K has the value,

$$\beta^{-1}(300K) = k_B \cdot 300K = 25\,meV = (1/40)\,eV \tag{73}$$

5. Phase space evolution and H-Theorem

Systems of *N* real particles occupy domains in *6N*-dimensional ***phase space***, rather than cells of a CA. Phase space is a product space described by continuous *3N* spatial $\{\vec{q}_i,\ i=1,...,N\}$ and *3N* momentum $\{\vec{p}_i,\ i=1,...,N\}$ coordinates. Therefore, the probabilities p_i of discrete cells *i* discussed previously is replaced by continuous, time dependent (*t*) distribution functions $\{f(\vec{q}_i,\vec{p}_i,t),\ i=1,...,N\}$ for the *N* particles. These functions are probability densities normalized to unity when integrated over the entire phase space,

$$\int d^3\vec{q}_i \int d^3 f(\vec{q}_i,\vec{p}_i,t) \equiv 1 \qquad \{i=1,...,N\} \tag{74}$$

Following the same line of arguments as before, the time dependent information content of an occupied multi-particle state is contained in the ***Boltzmann H-function*** (*eta*-function)

$$H(t) := \sum_{i=1}^{N} \int d^3\vec{q}_i \int d^3\vec{p}_i \ \{f(\vec{q}_i,\vec{p}_i,t)\cdot \ell n f(\vec{q}_i,\vec{p}_i,t)\} \le 0 \tag{75}$$

The *H* function is obviously equivalent to the negative of the information *S* given by the statistical entropy (cf. Equ.(25)). It is negative since the distribution functions are probability densities.

Based on very general principles, predictions can be made as to the spontaneous time evolution of the *H* function, or the equivalent statistical entropy function *S.* In the following, the entropy *S(t)* is expressed as

$$S(t) = -H(t) = -k_B \sum_{n=1}^{\Omega} p_n(t)\, \ell n\left(p_n(t)\right) \ge 0 \qquad \sum_{n=1}^{\Omega} p_n(t) \equiv 1 \tag{76}$$

in terms of time dependent (normalized) probabilities for discrete system states numbered by *n*.

This time dependence of the entropy function reflects an underlying dynamics, a transport process, which tends to redistribute the importance (or population) of the microscopic states and all of its attributes. The trend is equivalent to an entropy flux or current

$$j_s := \frac{dS}{dt} \tag{77}$$

If j_s has a finite magnitude, it defines a direction of increasing or decreasing diversity or spread in *a priori probabilities.*

The *a priori* probabilities p_n can be regarded as populations of these states which can be queried in experimental observations. If these populations are time dependent, there have to be microscopic transition probabilities w_{nm} connecting any state *n* and *m*. The transition probabilities describe the rate of change in the population of state *n* due to gain and loss from and to state *m* according to a balance "Master Equation,"

$$\frac{dp_n(t)}{dt} = \sum_m \left\{ \underbrace{w_{mn} \cdot p_m(t)}_{Gain} - \underbrace{w_{nm} \cdot p_n(t)}_{Loss} \right\} \tag{78}$$

For microscopic, quantal reasons, the transition probabilities are symmetric, $w_{nm} = w_{mn}$, which ensures time reversal invariance (detailed balance). Obviously, the Master Equation (78) is a classical approximation in that it neglects quantal interference terms involving transition amplitudes, rather than probabilities.

Now, the time derivative of the entropy function in Equ. (76), the entropy flux (Equ. (77)), can be calculated:

$$\frac{dS(t)}{dt} = -k_B \sum_{n=1}^{\Omega} \left\{ \left(\frac{dp_n(t)}{dt} \right) \ell n\left(p_n(t)\right) + p_n(t) \left(\frac{d\ell n\, p_n(t)}{dt} \right) \right\}; \qquad \frac{d}{dt} \sum_{n=1}^{\Omega} p_n(t) \equiv 0 \tag{79}$$

Evaluating the derivatives one obtains

$$\frac{dS(t)}{dt} = -k_B \sum_{n=1}^{\Omega} \left\{ \left(\frac{dp_n(t)}{dt} \right) \ell n\left(p_n(t)\right) + p_n(t) \left(\frac{1}{p_n(t)} \frac{dp_n(t)}{dt} \right) \right\} =$$
$$\frac{dS(t)}{dt} = -k_B \sum_{n=1}^{\Omega} \left(\frac{dp_n(t)}{dt} \right) \ell n\left(p_n(t)\right) - k_B \underbrace{\sum_n \frac{dp_n(t)}{dt}}_{=0} \qquad (80)$$

The last term drops out because of the conservation of total probability implied by Equ. (79). Now, inserting for dp_n/dt the expression given by the Master Equation (78), the second row in (80) reads,

$$\frac{dS(t)}{dt} = -k_B \sum_{n,m=1}^{\Omega} w_{mn} \cdot \left\{ p_m(t) - p_n(t) \right\} \ell n\left(p_n(t)\right) \qquad (81)$$

Here, use has been made of the symmetry of the transition probabilities w_{mn}. Since the two indices n and m run over the same range, this expression can also be written as,

$$\frac{dS(t)}{dt} = -k_B \sum_{n,m=1}^{\Omega} w_{mn} \left\{ p_n(t) - p_m(t) \right\} \ell n\left(p_m(t)\right) \qquad (82)$$

Taking the average of Equs. (81) and (82), a more symmetric expression is obtained fro the time rate of change of the entropy function:

$$\frac{dS(t)}{dt} = \frac{k_B}{2} \sum_{n,m=1}^{\Omega} w_{mn} \left\{ p_n(t) - p_m(t) \right\} \left[\ell n\left(p_n(t)\right) - \ell n\left(p_m(t)\right) \right] \qquad (83)$$

However, since $d\ell n(p)/dp > 0$, all terms in the sum are non-negative and therefore,

$$j_s = \frac{dS(t)}{dt} = -\frac{dH(t)}{dt} \geq 0 \qquad (84)$$

According to this derivation, the entropy S increases and the H function decreases in time, as long as the transition probabilities are finite, $w_{nm} = w_{mn} > 0$. The larger the differences between the populations p_i of different states are, the higher is the rate of entropy changes. When

$$p_n \approx const.; \quad n = 1,.....,\Omega \tag{85}$$

the S (or H) functions no longer change. The system described by such function has reached its asymptotic stationary state, also known as equilibrium state. ***This equilibrium state is characterized by maximum entropy corresponding to equal a priori probabilities p_n and chaotic dynamics***. While for a given theoretic model the expectation values of the functions S and H can be calculated exactly, there are also higher moments (fluctuations) to consider, since they depend on stochastic parameters, the probabilities p_n.

6. Gibbs stability criterion for random states

The situation of maximum entropy, where all accessible states are uniform and have equal *a priori* probabilities, is called "***equilibrium***." It will be shown further below how these information/entropy functions change in complex dynamical processes.

All systems where accessible states are not uniformly populated are in states of disequilibrium and have statistical entropies less than the maximum possible:

The equilibrium state is therefore defined by the variational condition

$$S(\vec{q}_i,\vec{p}_i) = S_{max}(\vec{q}_i,\vec{p}_i) =: S_{equ} \;\rightarrow\; \delta S(\vec{q}_i,\vec{p}_i) = 0 \qquad (86)$$

Here, δ stands for a variation with respect to the individual probability densities. Once a multi-particle system is in such an equilibrium state of maximum entropy, there is conceptionally ***no net driving force*** that would force it out of this state in one direction or another. However, such an equilibrium state can be either stable or unstable. Microscopically, there are always quantal fluctuations in all coordinates. Even systems presumably at rest show "zero-point fluctuations." In addition, physical particles move even classically from phase space cell to phase space cell, changing individual occupation probabilities (p_i or $f(\vec{q}_i,\vec{p}_i,t)$) instantaneously away from their respective equilibrium values. The magnitude of these fluctuations depend on their origin in classical or quantum dynamics. They may vary in size and follow a distribution in time or frequency (chance of occurrence). Therefore, the actual entropy at a given instant will reflect these fluctuations.

Connecting to discussions of stability in previous sections, one can obtain a stability criterion by studying the expansion of the entropy S of an actual system state about the equilibrium state ($\delta S = 0$),

$$S = S_{equ} + \delta S + \frac{1}{2}\delta^2 S + ... \approx S_{equ} + \frac{1}{2}\delta^2 S \qquad (87)$$

From this relation it simply follows that the state of maximum entropy is stable, only if fluctuations away from this state reduce the entropy,

$$\delta^2 S < 0 \tag{88}$$

This "Gibbs" stability criterion has to be applied in specific cases to identify the stable equilibrium. Stable equilibrium states are attractors of complex system, as will be demonstrated in later sections.

7. Specific probability distributions

See tutorial