Chapter 2
Learning with Boltzmann–Gibbs Statistical Mechanics

Πᾶν μέτρον ἄριστον [78]
Kleoboulos of Lindos (6th century B.C.)

2.1 Boltzmann–Gibbs Entropy

2.1.1 Entropic Forms

The entropic forms (1.1) and (1.3) that we have introduced in Chapter 1 correspond to the case where the (microscopic) states of the system are *discrete*. There are, however, cases in which the appropriate variables are *continuous*. For these, the BG entropy takes the form

$$S_{BG} = -k \int dx\; p(x)\; \ln[\sigma p(x)]\,, \tag{2.1}$$

with

$$\int dx\; p(x) = 1\,, \tag{2.2}$$

where $x/\sigma \in \mathbb{R}^D$, $D \geq 1$ being the dimension of the full space of microscopic states (called *Gibbs* Γ *phase-space* for classical Hamiltonian systems). Typically x carries physical units. The constant σ carries the same physical units as x, so that x/σ is a dimensionless quantity (we adopt from now on the notation $[x] = [\sigma]$, hence $[x/\sigma] = 1$). For example, if we are dealing with an isolated classical N-body Hamiltonian system of point masses interacting among them in d dimensions, we may use $\sigma = \hbar^{Nd}$. This standard choice comes of course from the fact that, at a sufficiently small scale, Newtonian mechanics becomes incorrect and we must rely on quantum mechanics. In this case, $D = 2dN$, where each of the d pairs of components of momentum and position of each of the N particles has been taken

C. Tsallis, *Introduction to Nonextensive Statistical Mechanics*,
DOI 10.1007/978-0-387-85359-8_2, © Springer Science+Business Media, LLC 2009

into account (we recall that $[momentum][position] = [\hbar]$). For the case of *equal probabilities* (i.e., $p(x) = 1/\Omega$, where Ω is the hypervolume of the admissible D-dimensional space), we have

$$S_{BG} = k \ln(\Omega/\sigma) . \tag{2.3}$$

A particular case of $p(x)$ is the following one:

$$p(x) = \sum_{i=1}^{W} p_i \, \Delta(x - x_i) \quad (W \equiv \Omega/\sigma) , \tag{2.4}$$

where $\Delta(x-x_i)$ denotes a normalized uniform distribution centered on x_i and whose "width" is σ (hence its height is $1/\sigma$). In this case, Eqs. (2.1), (2.2) and (2.3) precisely recover Eqs. (1.1), (1.2) and (1.3).

In both discrete and continuous cases that we have addressed until now, we were considering classical systems in the sense that all physical observables are real quantities and *not operators*. However, for intrinsically quantum systems, we must generalize the concept. In that case, the BG entropic form is to be written (as first introduced by von Neumann) in the following manner:

$$S_{BG} = -k \, Tr\rho \ln \rho , \tag{2.5}$$

with

$$Tr\rho = 1 , \tag{2.6}$$

where ρ is the *density matrix* acting on a W-dimensional Hilbert vectorial space (typically associated with the solutions of the Schroedinger equation with the chosen boundary conditions; in fact, quite frequently we have $W \to \infty$).

A particular case of ρ is when it is *diagonal*, i.e., the following one:

$$\rho_{ij} = p_i \, \delta_{ij} , \tag{2.7}$$

where δ_{ij} denotes Kroenecker's delta function. In this case, Eqs. (2.5) and (2.6) exactly recover Eqs. (1.1) and (1.2).

All three entropic forms (1.1), (2.1), and (2.5) will be generically referred in the present book as *BG-entropy* because they are all based on a logarithmic measure for *disorder*. Although we shall use one or the other for specific purposes, we shall mainly address the simple form expressed in Eq. (1.1).

2.1.2 Properties

2.1.2.1 Non-negativity

If we know with *certainty* the state of the system, then $p_{i_0} = 1$, and $p_i = 0, \forall i \neq i_0$. Then it follows that $S_{BG} = 0$, where we have taken into account that $\lim_{x\to 0}(x \ln x) = 0$. In any other case, we have $p_i < 1$ for at least two different values of i. We can therefore write Eq. (1.1) as follows:

$$S_{BG} = -k\langle \ln p_i \rangle = k\left\langle \ln \frac{1}{p_i} \right\rangle , \tag{2.8}$$

where $\langle \cdots \rangle \equiv \sum_{i=1}^{W} p_i(\ldots)$ is the standard *mean value*. Since $\ln(1/p_i) > 0\,(\forall i)$, it clearly follows that S_{BG} is *positive*.

2.1.2.2 Maximal at Equal Probabilities

Energy is a concept which definitively takes into account the physical nature of the system. *Less so, in some sense, the BG entropy.*[1] This entropy depends of course on the total number of possible microscopic configurations of the system, but it is insensitive to its specific physical support; it only takes into account the (abstract) probabilistic information on the system. Let us make a trivial illustration: a spin that can be up or down (with regard to some external magnetic field), a coin that comes head or tail, and a computer bit which can be 0 or 1 are all equivalent for the concept of entropy. Consequently, entropy is expected to be a functional which is *invariant with regard to any permutations of the states*. Indeed, expression (1.1) exhibits this invariance through the form of a sum. Consequently, if $W > 1$, the entropy must have an extremum (maximum or minimum), and this must occur for equal probabilities. Indeed, this is the unique possibility for which the entropy is invariant with regard to the permutation of *any two* states. It is easy to verify that it is a maximum, and not a minimum. In fact, the identification as a maximum (and not a minimum) will become obvious when we shall prove, later on, that S_{BG} is a *concave* functional. Of course, the expression that S_{BG} takes for equal probabilities has already been indicated in Eq. (1.3).

2.1.2.3 Expansibility

Adding to a system new possible states with *zero* probability should not modify the entropy. This is precisely what is satisfied by S_{BG} if we take into account the

---

[1] This statement is to be revisited for the more general entropy S_q. Indeed, as we shall see, the index q does depend on some universal aspects of the physical system, e.g., the type of inflexion of a dissipative unimodal map, or, possibly, the type of power-law decay of long-range interactions for Hamiltonian systems.

already-mentioned property $\lim_{x\to 0}(x \ln x) = 0$. So, we have that

$$S_{BG}(p_1, p_2, \ldots, p_W, 0) = S_{BG}(p_1, p_2, \ldots, p_W)\,. \tag{2.9}$$

2.1.2.4 Additivity

Let $\mathcal{O}$ be a physical quantity associated with a given system, and let A and B be two probabilistically independent subsystems. We shall use the term *additive* if and only if $\mathcal{O}(A+B) = \mathcal{O}(A) + \mathcal{O}(B)$. If so, it is clear that if we have N equal systems, then $\mathcal{O}(N) = N\mathcal{O}(1)$, where the notation is self-explanatory. A weaker condition is $\mathcal{O}(N) \sim N\Omega$ for $N \to \infty$, with $0 < |\Omega| < \infty$, i.e., $\lim_{N\to\infty} \mathcal{O}(N)/N$ is finite (generically $\Omega \neq \mathcal{O}(1)$). In this case, the expression *asymptotically additive* might be used. Clearly, any observable, which is additive with regard to a given composition law, is asymptotically additive (with $\Omega = \mathcal{O}(1)$), but the opposite is not necessarily true.

It is straightforwardly verified that, if A and B are *independent*, i.e., if the *joint probability* satisfies $p_{ij}^{A+B} = p_i^A p_j^B$ $(\forall (ij))$, then

$$S_{BG}(A+B) = S_{BG}(A) + S_{BG}(B)\,. \tag{2.10}$$

Therefore, the entropy S_{BG} is *additive*.

2.1.2.5 Concavity

Let us assume two arbitrary and different probability sets, namely $\{p_i\}$ and $\{p'_i\}$, associated with a single system having W states. We define an *intermediate* probability set as follows:

$$p''_i = \lambda p_i + (1-\lambda)p'_i \quad (\forall i;\, 0 < \lambda < 1)\,. \tag{2.11}$$

The functional $S_{BG}(\{p_i\})$ (or any other functional in fact) is said *concave if and only if*

$$S_{BG}(\{p''_i\}) > \lambda S_{BG}(\{p_i\}) + (1-\lambda) S_{BG}(\{p'_i\})\,. \tag{2.12}$$

This is indeed satisfied by S_{BG}. The proof is straightforward. Because of its *negative second derivative*, the (continuous) function $-x \ln x$ $(x > 0)$ satisfies

$$-p''_i \ln p''_i > \lambda(-p_i \ln p_i) + (1-\lambda)(-p'_i \ln p'_i) \quad (\forall i;\, 0 < \lambda < 1)\,. \tag{2.13}$$

Applying $\sum_{i=1}^{W}$ on both sides of this inequality, we immediately obtain Eq. (2.12), i.e., the *concavity* of S_{BG}.

2.1.2.6 Lesche-Stability or Experimental Robustness

An entropic form $S(\{p_i\})$ (or any other functional of the probabilities, in fact) is said *Lesche-stable* or *experimentally robust* [79] if and only if it satisfies the following continuity property. Two probability distributions $\{p_i\}$ and $\{p'_i\}$ are said *close* if they satisfy the metric property:

$$D \equiv \sum_{i=1}^{W} |p_i - p'_i| \le d_\epsilon \,, \tag{2.14}$$

where d_ϵ is a small real number. Then, experimental robustness is verified if, for any $\epsilon > 0$, a d_ϵ exists such that $D \le d_\epsilon$ implies

$$R \equiv \left| \frac{S(\{p_i\}) - S(\{p'_i\})}{S_{max}} \right| < \epsilon \,, \tag{2.15}$$

where S_{max} is the maximal value that the entropic form can achieve (assuming its extremum corresponds to a maximum and not a minimum). For S_{BG} the maximal value is of course $\ln W$.

Condition (2.15) should be satisfied under all possible situations, including for $W \to \infty$. This implies that the condition

$$\lim_{d_\epsilon \to 0} \lim_{W \to \infty} \left| \frac{S(\{p_i\}) - S(\{p'_i\})}{S_{max}} \right| = 0 \tag{2.16}$$

should *also* be satisfied, in addition to $\lim_{W\to\infty} \lim_{d\to 0} \left| \frac{S(\{p_i\}) - S(\{p'_i\})}{S_{max}} \right| = 0$, which is of course always satisfied.

What this property essentially guarantees is that *similar* experiments performed onto *similar* physical systems should provide *similar* results (i.e., a small percentage discrepancy) for the measured physical functionals. Lesche showed [79] that S_{BG} is experimentally robust, whereas the Renyi entropy $S_q^R \equiv \frac{\ln \sum_{i=1}^{W} p_i^q}{1-q}$ is not. See Fig. 2.1.

It is in principle possible to use, as a concept for *distance*, a quantity different from that used in Eq. (2.14). We could use for instance the following definition:

$$D_\mu \equiv \left[\sum_{i=1}^{W} |p_i - p'_i|^\mu \right]^{1/\mu} \quad (\mu > 0) \,. \tag{2.17}$$

Equation (2.14) corresponds to $\mu = 1$. The Pythagorean metric corresponds to $\mu = 2$. What about other values of μ? It happens that only for $\mu \ge 1$ the triangular inequality is satisfied, and consequently it does constitute a metric. Still, why not

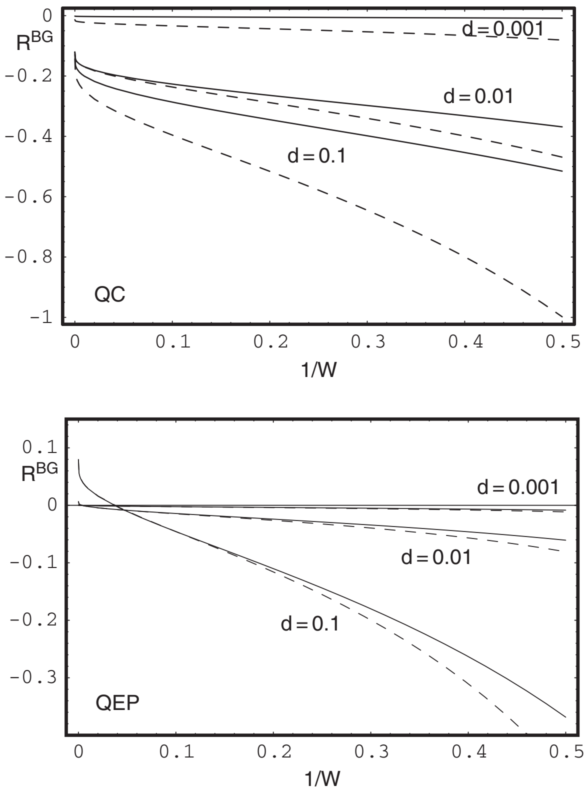

Fig. 2.1 Illustration of the Lesche-stability of S_{BG}. QC and QEP stand for *quasi-certainty* and *quasi-equal-probabilities*, respectively (see details in [110, 113]).

using values of $\mu > 1$? Because, only for $\mu = 1$, the distance D does *not* depend on W, which makes it appropriate for a generic property [80].

We come back in Section 3.2.2 onto this interesting property introduced by Lesche.

2.1.2.7 Shannon Uniqueness Theorem

Let us assume that an entropic form $S(\{p_i\})$ satisfies the following properties:

$$(i)\;\; S(\{p_i\})\, is\ a\ continuous\ function\ of\ \{p_i\}; \tag{2.18}$$

$$(ii)\;\; S(p_i = 1/W, \forall i)\, monotonically\ increases\ with\ the\ total\ number\ of\ possibilities\ W; \tag{2.19}$$

$$(iii)\; S(A + B) = S(A) + S(B) \quad if\ p_{ij}^{A+B} = p_i^A p_j^B\, \forall (i, j)\,, \tag{2.20}$$

$$where\ S(A+B) \equiv S(\{p_{ij}^{A+B}\}),\ S(A) \equiv S(\{p_i^A\})\ (p_i^A \equiv \sum_{j=1}^{W_B} p_{ij}^{A+B}),$$

$$and \quad S(B) \equiv S(\{p_j^B\})\ (p_j^B \equiv \sum_{i=1}^{W_A} p_{ij}^{A+B});$$

$$(iv)\ S(\{p_i\}) = S(p_L, p_M) + p_L S(\{p_i/p_L\}) + p_M S(\{p_i/p_M\}) \tag{2.21}$$

$$with \quad p_L \equiv \sum_{L\,terms} p_i\ ,\ p_M \equiv \sum_{M\,terms} p_i\ ,$$

$$L + M = W\ ,\ and\ p_L + p_M = 1\ .$$

Then and only then [25]

$$S(\{p_i\}) = -k \sum_{i=1}^{W} p_i \ln p_i \quad (k > 0)\,. \tag{2.22}$$

2.1.2.8 Khinchin Uniqueness Theorem

Let us assume that an entropic form $S(\{p_i\})$ satisfies the following properties:

$$(i)\ \ S(\{p_i\})\ is\ a\ continuous\ function\ of\ \{p_i\}; \tag{2.23}$$

$$(ii)\ \ S(p_i = 1/W, \forall i)\ monotonically\ increases\ with\ the\ total\ number\ of\ possibilities\ W; \tag{2.24}$$

$$(iii)\ S(p_1, p_2, \ldots, p_W, 0) = S(p_1, p_2, \ldots, p_W); \tag{2.25}$$

$$(iv)\ \ S(A+B) = S(A) + S(B|A), \tag{2.26}$$

$$where\ S(A+B) \equiv S(\{p_{ij}^{A+B}\}),\ S(A) \equiv S(\{p_i^A\})\ (p_i^A \equiv \sum_{j=1}^{W_B} p_{ij}^{A+B}),$$

$$and\ the\ conditional\ entropy\ S(B|A) \equiv \sum_{i=1}^{W_A} p_i^A S(\{p_{ij}^{A+B}/p_i^A\})\,.$$

Then and only then [81]

$$S(\{p_i\}) = -k \sum_{i=1}^{W} p_i \ln p_i\ (k > 0)\,. \tag{2.27}$$

2.1.2.9 Composability

A dimensionless entropic form $S(\{p_i\})$ (i.e., whenever expressed in appropriate conventional units, e.g., in units of k) is said *composable* if the entropy $S(A + B)$ corresponding to a system composed of two *independent* subsystems A and B can be expressed in the form

$$S(A+B) = F(S(A), S(B); \{\eta\}) \,, \tag{2.28}$$

where $F(x, y; \{\eta\})$ is a function which, besides depending symmetrically on (x, y), depends on a (typically small) set of universal indices $\{\eta\}$. In other words, it does *not* depend on the microscopic configurations of A and B. Equivalently, we are able to macroscopically calculate the entropy of the composed system without any need of entering into the knowledge of the microscopic states of the subsystems. This property appears to be a natural one for an entropic form if we desire to use it as a basis for a statistical mechanics which would naturally connect to thermodynamics.

The BG entropy is composable since it satisfies Eq. (2.10). In other words, we have $F(x, y) = x + y$. Since S_{BG} is nonparametric, no index exists.

2.1.2.10 Sensitivity to the Initial Conditions, Entropy Production per Unit Time, and a Pesin-Like Identity

For a one-dimensional dynamical system (characterized by the variable x) the sensitivity to the initial conditions ξ is defined as follows:

$$\xi \equiv \lim_{\Delta x(0)\to 0} \frac{\Delta x(t)}{\Delta x(0)} \,. \tag{2.29}$$

It can be shown [82, 83] that ξ paradigmatically satisfies the equation

$$\frac{d\xi}{dt} = \lambda_1 \, \xi \,, \tag{2.30}$$

whose solution is given by

$$\xi = e^{\lambda_1 t} \,. \tag{2.31}$$

(The meaning of the subscript 1 will become transparent later on). If the *Lyapunov exponent* $\lambda_1 > 0$ ($\lambda_1 < 0$), the system will be said to be *strongly chaotic* (*regular*). The case where $\lambda_1 = 0$ is sometimes called *marginal* and will be extensively addressed later on.

At this point let us briefly review, without proof, some basic notions of nonlinear dynamical systems. If the system is d-dimensional (i.e., it evolves in a phase-space whose d-dimensional Lebesgue measure is finite), it has d Lyapunov exponents: d_+ of them are positive, d_- are negative, and d_0 vanish, hence $d_+ + d_- + d_0 = d$. Let us order them all from the largest to the smallest: $\lambda_1^{(1)} \ge \lambda_1^{(2)} \ge \ldots \ge \lambda_1^{(d_+)} > \lambda_1^{(d_+ +1)} = \lambda_1^{(d_+ +2)} = \ldots = \lambda_1^{(d_+ + d_0)} = 0 > \lambda_1^{(d_+ + d_0 + 1)} \ge \lambda_1^{(d_+ + d_0 + 2)} \ge \ldots \ge \lambda_1^{(d)}$. An infinitely small *segment* (having then a well defined one-dimensional Lebesgue measure) diverges like $e^{\lambda_1^{(1)} t}$; this precisely is the case focused in Eq. (2.31). An infinitely small *area* (having then a well defined two-dimensional Lebesgue measure) diverges like $e^{(\lambda_1^{(1)} + \lambda_1^{(2)})t}$. An infinitely small volume diverges like $e^{(\lambda_1^{(1)} + \lambda_1^{(2)} + \lambda_1^{(3)})t}$, and so on. An infinitely small d-dimensional hypervolume evolves like $e^{[\sum_{r=1}^{d} \lambda_1^{(r)}]t}$. If the system is *conservative*, i.e., if the infinitely small d-dimensional hypervolume remains constant with time, then it follows that $\sum_{r=1}^{d} \lambda_1^{(r)} = 0$. An important particular class of conservative systems is constituted by the so-called *symplectic* ones. For these, d is an even integer, and the Lyapunov exponents are coupled two by two as follows: $\lambda_1^{(1)} = -\lambda_1^{(d)} \geq \lambda_1^{(2)} = -\lambda_1^{(d-1)} \geq \ldots \geq \lambda_1^{(d_+)} = -\lambda_1^{(d_++d_0+1)} \geq \lambda_1^{(d_++1)} = \ldots = \lambda_1^{(d_++d_0)} = 0$. Consistently, such systems have $d_+ = d_-$ and d_0 is an even integer. The most popular illustration of symplectic systems is the Hamiltonian systems. They are conservative, which precisely is what the classical Liouville theorem states!

Do all these degrees of freedom contribute, as time evolves, to the erratic exploration of the phase-space? *No, they do not*. Only those associated with the d_+ positive Lyapunov exponents, and some of the d_0 vanishing ones, do. Consistently, it is only these which contribute to our loss of information, as time evolves, about the location in phase-space of a set of initial conditions. As we shall see, these remarks enable an intuitive understanding to the so-called Pesin identity, that we will soon state.

Let us now address the interesting question of the BG entropy production as time t increases. More than one entropy production can be defined as a function of time. Two basic choices are the so-called *Kolmogorov–Sinai entropy* (or *KS entropy rate*) [84], based on a single trajectory in phase-space, and the one associated to the evolution of an *ensemble of initial conditions*. We shall preferentially use here the latter, because of its sensibly higher computational tractability. In fact, excepting for pathological cases, they both coincide.

Let us schematically describe the *Kolmogorov–Sinai entropy rate* concept or *metric entropy* [83, 84, 286]. We first partition the phase-space in two regions, noted A and B. Then we choose a generic initial condition (the final result will not depend on this choice) and, applying the specific dynamics of the system at equal and finite time intervals τ, we generate a long string (infinitely long in principle), say $ABBBAABBABAAA\ldots$. Then we analyze words of length $l = 1$. In our case, there are only two such words, namely A and B. The analysis consists in running along the string a window whose width is l, and determining the probabilities p_A and p_B of the words A and B, respectively. Finally, we calculate the entropy $S_{BG}(l = 1) = -p_A \ln p_A - p_B \ln p_B$. Then we repeat for words whose length equals $l = 2$. In our case, there are four such words, namely AA, AB, BA, BB. Running along the string a $l = 2$ window letter by letter, we determine the probabilities $p_{AA}, p_{AB}, p_{BA}, p_{BB}$, hence the entropy $S_{BG}(l = 2) = -p_{AA} \ln p_{AA} - p_{AB} \ln p_{AB} - p_{BA} \ln p_{BA} - p_{BB} \ln p_{BB}$. Then we repeat for $l = 3, 4, \ldots$ and calculate the corresponding values for $S_{BG}(l)$. We then choose another two-partition, say A' and B', and repeat the whole procedure. Then we do in principle for all possible two partitions. Then we go to three partitions, i.e., the alphabet will be now constituted by three letters, say A, B, and C. We repeat the previous procedure for $l = 1$ (corresponding to the words A, B, C), then for $l = 2$ (corresponding to the words $AA, AB, AC, BA, BB, BC, CA, CB, CC$), etc. Then we run windows with $l = 3, 4, \ldots$. We then consider a different three-partition, say A', B', and $C' \ldots$ Then we consider four-partitions, and so on. Of all these entropies we retain the *supremum*. In the appropriate limits of infinitely fine partitions and

$\tau \to 0$ we obtain finally the largest rate of increase of the BG entropy. This is basically the Kolmogorov–Sinai entropy rate.

It is not necessary to insist on how deeply inconvenient this definition can be for any computational implementation! Fortunately, a different type of entropy production can be defined [85], whose computational implementation is usually very simple. It is defined as follows. First partition the phase-space into W little cells (normally equal in size) and denote them with $i = 1, 2, \ldots, W$. Then randomly place M initial conditions in one of those W cells (if $d_+ \geq 1$, normally the result will not depend on this choice). And then follow, as time evolves, the number of points $M_i(t)$ in each cell ($\sum_{i=1}^{W} M_i(t) = M$). Define the probability set $p_i(t) \equiv M_i(t)/M \quad (\forall i)$, and calculate finally $S_{BG}(t)$ through Eq. (1.1). The *entropy production per unit time* is defined as

$$K_1 \equiv \lim_{t\to\infty} \lim_{W\to\infty} \lim_{M\to\infty} \frac{S_{BG}(t)}{t} \,. \tag{2.32}$$

The Pesin identity [86], or more precisely the Pesin-like identity that we shall use here, states, for large classes of dynamical systems [85],

$$K_1 = \sum_{r=1}^{d_+} \lambda_1^{(r)} \,. \tag{2.33}$$

As it will become gradually clear along the book, this relationship (and its q-generalization) will play an important role in the determination of the particular entropic form which is adequate for a given nonlinear dynamical system.

2.2 Kullback–Leibler Relative Entropy

In many problems the question arises on how different are two probability distributions p and $p^{(0)}$; for reasons that will become clear soon, $p^{(0)}$ will be referred to as the *reference*. It becomes therefore interesting to define some sort of "distance" between them. One possibility is of course the distance introduced in Eq. (2.17). In other words, for say continuous distributions, we can use

$$D_\mu(p, p^{(0)}) \equiv \left[\int dx\, |p(x) - p^{(0)}(x)|^\mu\right]^{1/\mu} \quad (\mu > 0) \,. \tag{2.34}$$

In general we have that $D_\mu(p, p^{(0)}) = D_\mu(p^{(0)}, p)$, and that $D_\mu(p, p^{(0)}) = 0$ if and only if $p(x) = p^{(0)}(x)$ almost everywhere. We remind, however, that the triangular inequality is satisfied only for $\mu \geq 1$. Therefore, only then the distance constitutes a metric. If $p(x) = \sum_{i=1}^{W} p_i\, \Delta(x - x_i)$ and $p^{(0)}(x) = \sum_{i=1}^{W} p_i^{(0)}\, \Delta(x - x_i)$, (see Eq. (2.4)) Eq. (2.34) leads to

$$D_\mu(p, p^{(0)}) \equiv \Big[\sum_{i=1}^{W} |p_i - p_i^{(0)}|^\mu\Big]^{1/\mu} \quad (\mu > 0)\,, \tag{2.35}$$

which exactly recovers Eq. (2.17).

For some purposes, this definition of distance is quite convenient. For others, the *Kullback–Leibler entropy* [87] has been introduced (see, for instance, [88, 92] and references therein). It is occasionally called *cross entropy*, or *relative entropy*, or *mutual information*, and it is defined as follows:

$$I_1(p, p^{(0)}) \equiv \int dx\; p(x) \ln\Big[\frac{p(x)}{p^{(0)}(x)}\Big] = -\int dx\; p(x) \ln\Big[\frac{p^{(0)}(x)}{p(x)}\Big]. \tag{2.36}$$

It can be proved, by using $\ln r \geq 1 - 1/r$ (with $r \equiv p(x)/p^{(0)}(x) > 0$), that $I_1(p, p^{(0)}) \geq 0$, the equality being valid if and only if $p(x) = p^{(0)}(x)$ almost everywhere. It is clear that in general $I_1(p, p^{(0)}) \neq I_1(p^{(0)}, p)$. This inconvenience is sometimes overcome by using the symmetrized quantity $[I_1(p, p^{(0)}) + I_1(p^{(0)}, p)]/2$.

$I_1(p, p^{(0)})$ (like the distance (2.34)) has the property of being invariant under variable transformation. Indeed, if we make $x = f(y)$, the measure preservation implies $p(x)dx = \tilde{p}(y)dy$. Since $p(x)/p^{(0)}(x) = \tilde{p}(x)/\tilde{p}^{(0)}(x)$, we have $I_1(p, p^{(0)}) = I_1(\tilde{p}, \tilde{p}^{(0)})$, which proves the above-mentioned invariance. The BG entropy in its continuous (*not* in its discrete) form $S_{BG} = -\int dx\; p(x) \ln p(x)$ lacks this important property. Because of this fact, the BG entropy is advantageously replaced, in some calculations, by the Kullback–Leibler one. Depending on the particular problem, the referential distribution $p^{(0)}(x)$ is frequently taken to be a standard distribution such as the uniform, or Gaussian, or Lorentzian, or Poisson or BG ones. When $p^{(0)}(x)$ is chosen to be the uniform distribution on a compact support of Lebesgue measure W, we have the relation

$$I_1(p, 1/W) = \ln W - S_{BG}(p)\,. \tag{2.37}$$

Because of relations of this kind, the minimization of the Kulback–Leibler entropy is sometimes used instead of the maximization of the Boltzmann–Gibbs–Shannon entropy.

Although convenient for a variety of purposes, $I_1(p, p^{(0)})$ has a disadvantage. It is needed that $p(x)$ and $p^{(0)}(x)$ *simultaneously* vanish, if they do so for certain values of x (this property is usually referred to as being *absolutely continuous*). Indeed, it is evident that otherwise the quantity $I_1(p, p^{(0)})$ becomes ill-defined. To overcome this difficulty, a different distance has been defined along the lines of the Kullback–Leibler entropy. We refer to the so-called *Jensen–Shannon divergence*. Although interesting in many respects, its study would take us too far from our present line. Details can be seen in [93, 94] and references therein.

Let us mention also that, for discrete probabilities, definition (2.36) leads to

$$I_1(p, p^{(0)}) \equiv \sum_{i=1}^{W} p_i \ln\Big[\frac{p_i}{p_i^{(0)}}\Big] = -\sum_{i=1}^{W} p_i \ln\Big[\frac{p_i^{(0)}}{p_i}\Big]. \tag{2.38}$$

Various other interesting related properties can be found in [95, 96].

2.3 Constraints and Entropy Optimization

The most simple entropic optimization cases are those worked out in what follows.

2.3.1 Imposing the Mean Value of the Variable

In addition to

$$\int_0^\infty dx\; p(x) = 1\,, \tag{2.39}$$

we might know the mean value of the variable, i.e.,

$$\langle x \rangle \equiv \int_0^\infty dx\; x p(x) = X^{(1)}\,. \tag{2.40}$$

By using the Lagrange method to find the optimizing distribution, we define

$$\Phi[p] \equiv -\int_0^\infty dx\; p(x) \ln p(x) - \alpha \int_0^\infty dx\; p(x) - \beta^{(1)} \int_0^\infty dx\; x p(x)\,, \tag{2.41}$$

and then impose $\delta\Phi[p]/\delta p(x) = 0$. We straightforwardly obtain $1 + \ln p_{opt} + \alpha + \beta^{(1)} x = 0$ (*opt* stands for *optimal*), hence

$$p_{opt} = \frac{e^{-\beta^{(1)} x}}{\int_0^\infty dx\; e^{-\beta^{(1)} x}} = \beta^{(1)}\, e^{-\beta^{(1)} x}, \tag{2.42}$$

where we have used condition (2.39) to eliminate the Lagrange parameter α. By using condition (2.40), we obtain the following relation for the Lagrange parameter $\beta^{(1)}$:

$$\beta^{(1)} = \frac{1}{X^{(1)}}\,, \tag{2.43}$$

hence, replacing in (2.42),

$$p_{opt} = \frac{e^{-x/X^{(1)}}}{X^{(1)}} \,. \tag{2.44}$$

2.3.2 Imposing the Mean Value of the Squared Variable

Another simple and quite frequent case is when we know that $\langle x \rangle = 0$. In such case, in addition to

$$\int_{-\infty}^{\infty} dx \; p(x) = 1 \,, \tag{2.45}$$

we might know the mean value of the squared variable, i.e.,

$$\langle x^2 \rangle \equiv \int_{-\infty}^{\infty} dx \, x^2 p(x) = X^{(2)} > 0 \,. \tag{2.46}$$

By using, as before, the Lagrange method to find the optimizing distribution, we define

$$\Phi[p] \equiv -\int_{-\infty}^{\infty} dx \; p(x) \ln p(x) - \alpha \int_{-\infty}^{\infty} dx \; p(x) - \beta^{(2)} \int_{-\infty}^{\infty} dx \, x^2 p(x) \,, \tag{2.47}$$

and then impose $\delta\Phi[p]/\delta p(x) = 0$. We straightforwardly obtain $1 + \ln p_{opt} + \alpha + \beta^{(2)} x^2 = 0$, hence

$$p_{opt} = \frac{e^{-\beta^{(2)} x^2}}{\int_{-\infty}^{\infty} dx \; e^{-\beta^{(2)} x^2}} = \sqrt{\frac{\beta^{(2)}}{\pi}} \, e^{-\beta^{(2)} x^2} , \tag{2.48}$$

where we have used condition (2.45) to eliminate the Lagrange parameter α.

By using condition (2.46), we obtain the following relation for the Lagrange parameter $\beta^{(2)}$:

$$\beta^{(2)} = \frac{1}{2X^{(2)}} \,, \tag{2.49}$$

hence, replacing in (2.48),

$$p_{opt} = \frac{e^{-x^2/(2X^{(2)})}}{\sqrt{2\pi \, X^{(2)}}} \,. \tag{2.50}$$

We thus see the very basic connection between Gaussian distributions and BG entropy.

2.3.3 *Imposing the Mean Values of both the Variable and Its Square*

Let us unify here the two preceding subsections. We impose

$$\int_{-\infty}^{\infty} dx\ p(x) = 1 \tag{2.51}$$

and, in addition to this, we know that

$$\langle x \rangle \equiv \int_{-\infty}^{\infty} dx\ xp(x) = X^{(1)}\,, \tag{2.52}$$

and

$$\langle (x-\langle x \rangle)^2 \rangle \equiv \int_{-\infty}^{\infty} dx\ (x-\langle x \rangle)^2 p(x) = X^{(2)} - (X^{(1)})^2 \equiv M^{(2)} > 0\,. \tag{2.53}$$

By using once again the Lagrange method, we define

$$\begin{aligned} \Phi[p] \equiv &- \int_{-\infty}^{\infty} dx\ p(x) \ln p(x) - \alpha \int_{-\infty}^{\infty} dx\ p(x) \\ &- \beta^{(1)} \int_{-\infty}^{\infty} dx\ x\ p(x) - \beta^{(2)} \int_{-\infty}^{\infty} dx\ (x-\langle x \rangle)^2 p(x)\,, \end{aligned} \tag{2.54}$$

and then impose $\delta\Phi[p]/\delta p(x) = 0$. We straightforwardly obtain $1 + \ln p_{opt} + \alpha + \beta^{(1)} x + \beta^{(2)}(x - \langle x \rangle)^2 = 0$, hence

$$p_{opt} = \frac{e^{-\beta^{(1)}x - \beta^{(2)}(x-\langle x \rangle)^2}}{\int_{-\infty}^{\infty} dx\ e^{-\beta^{(1)}x - \beta^{(2)}(x-\langle x \rangle)^2}} = \sqrt{\frac{\beta^{(2)}}{\pi}}\ e^{-\beta^{(2)}(x-\langle x \rangle)^2}\,, \tag{2.55}$$

where we have used condition (2.51) to eliminate the Lagrange parameter α. By using conditions (2.52) and (2.53), we obtain the following relations for the Lagrange parameters $\beta^{(1)}$ and $\beta^{(2)}$:

$$\beta^{(1)} = \frac{1}{X^{(1)}}\,, \tag{2.56}$$

and

$$\beta^{(2)} = \frac{1}{2[X^{(2)} - (X^{(1)})^2]}\,. \tag{2.57}$$

Replacing (2.57) in (2.55), we finally obtain

$$p_{opt} = \frac{e^{-\frac{(x-X^{(1)})^2}{2[X^{(2)}-(X^{(1)})^2]}}}{\sqrt{2\pi[X^{(2)} - (X^{(1)})^2]}} \,. \tag{2.58}$$

We see that the only effect of a nonzero mean value of x is to re-center the Gaussian.

2.3.4 Others

A quite general situation would be to impose, in addition to

$$\int dx\, p(x) = 1 \,, \tag{2.59}$$

the constraint

$$\int dx\, f(x)\, p(x) = F \,, \tag{2.60}$$

where $f(x)$ is some known function and F a known number. We obtain

$$p_{opt} = \frac{e^{-\beta f(x)}}{\int dx\, e^{-\beta f(x)}} \,. \tag{2.61}$$

It is clear that, by appropriately choosing $f(x)$, we can force $p_{opt}(x)$ to be virtually *any* distribution we wish. For example, by choosing $f(x) = |x|^\gamma \quad (\gamma \in \mathbb{R})$, we obtain a generic stretched exponential $p_{opt}(x) \propto e^{-\beta|x|^\gamma}$; by choosing $f(x) = \ln x$, we obtain for $p_{opt}(x)$ a power law. But the use of such procedures hardly has any epistemological interest at all, since it provides no hint onto the underlying nature of the problem. Only choices such as $f(x) = x$ or $f(x) = x^2$ are sound since such constraints correspond to very generic informational features, namely the *location of the center* and the *width* of the distribution. Other choices are, unless some exceptional fact enters into consideration (e.g., $f(x)$ being a constant of motion of the system), quite ad hoc and uninteresting. Of course, this mathematical fact is by no means exclusive of S_{BG} : the same holds for virtually *any* entropic form.

2.4 Boltzmann–Gibbs Statistical Mechanics and Thermodynamics

There are many formal manners for deriving the BG entropy and its associated probability distribution for thermal equilibrium. *None of them uses exclusively first principle arguments*, i.e., arguments that entirely remain at the level of mechanics (classical, quantum, relativistic, or any other). That surely was, as previously mentioned, one of the central scientific goals that Boltzmann pursued his entire life, but, although he probably had a strong intuition about this point, he died without succeeding. The difficulties are so heavy that even today we do not know how to do this. At first sight, this might seem surprising given the fact that S_{BG} and the BG weight enjoy the universal acceptance that we all know. So, let us illustrate our statement more precisely. Assume that we have a quite generic many-body short-range-interacting Hamiltonian. We currently know that its thermal equilibrium is described by the BG weight. What we still do not know is *how* to derive this important result from purely mechanical and statistical logical steps, i.e., without using a priori generic dynamical hypothesis such as *ergodicity*, or a priori postulating the validity of macroscopic relations such as some or all of the principles of thermodynamics. For example, Fisher et al. [97–99] proved long ago, for a vast class of short-range-interacting Hamiltonians, that the thermal equilibrium physical quantities are computable within standard BG statistical mechanics. Such a proof, no matter how precious might it be, does *not* prove also that this statistics *indeed* provides the correct description at thermal equilibrium. Rephrasing, it proves that BG statistics *can* be the correct one, but it does *not* prove that it *is* the correct one. Clearly, there is no reasonable doubt today that, for such systems, BG is the correct one. It is nevertheless instructive that the logical implications of the available proofs be outlined.

On a similar vein, even for the case of long-range-interacting Hamiltonians (e.g., infinitely-long-range interactions), the standard BG calculations can still be performed through convenient renormalizations of the coupling constants (e.g., *a la* Kac, or through the usual mean field approximation recipe of artificially dividing the coupling constant by the number N of particles raised to some appropriate power). The possibility of computability does *by no means* prove, strictly speaking, that BG statistics is the correct description. And certainly it does not enlighten us on what the necessary and sufficient *first-principle* conditions could be for the BG description to be the adequate one.

In spite of all these mathematical difficulties, at least one nontrivial example has been advanced in the literature [100] for which it has been possible to exhibit *numerically* the BG weight by *exclusively* using Newton's $\mathbf{F} = m\mathbf{a}$ as microscopic dynamics, *with no thermostatistical assumption of any kind*.

Let us anticipate that these and worse difficulties exist for the considerably more subtle situations that will be addressed in nonextensive statistical mechanics.

In what follows, we shall conform to more traditional, though epistemologically less ambitious, paths. We shall primarily follow the Gibbs' elegant lines of first *postulating* an entropic form, and then using it, *without proof*, as the basis for a variational principle including appropriate constraints. The philosophy of such path is quite clear. It is a form of *Occam' s razor*, where we use all that we know and not more than we know. This is obviously extremely attractive from a conceptual standpoint. However, that its mathematical implementation is to be done with a *given specific entropic functional* with *given specific constraints* is of course far from trivial! After 130 years of impressive success, there can be no doubt that BG

concepts and statistical mechanics provide the correct connection between microscopic and macroscopic laws for a vast class of physical systems. But – we insist – the mathematically *precise* qualification of this class remains an open question.

2.4.1 Isolated System – Microcanonical Ensemble

In this and subsequent subsections, we briefly review BG statistical mechanics (see, for instance, [35]). We consider a quantum Hamiltonian system constituted by N interacting particles under specific boundary conditions, and denote by $\{E_i\}$ its energy eigenvalues.

The microcanonical ensemble corresponds to an isolated N-particle system whose total energy U is known within some precision δU (to be in fact taken at its zero limit at the appropriate mathematical stage). The number of states i with $U \le E_i \le U + \delta U$ is denoted by W. *Assuming* that the system is such that its dynamics leads to ergodicity at its stationary state (thermal equilibrium), we assume that all such states are equally probable, i.e., $p_i = 1/W$, and the entropy is given by Eq. (1.3). The temperature T is introduced through

$$\frac{1}{T} \equiv \frac{\partial S_{BG}}{\partial U} = k\frac{\partial \ln W}{\partial U} \,. \tag{2.62}$$

2.4.2 In the Presence of a Thermostat – Canonical Ensemble

The canonical ensemble corresponds to an N-particle system defined in a Hilbert space whose dimension is noted W, and which is in longstanding thermal contact with a (infinitely large) thermostat at temperature T. Its exact energy is unknown, but its mean energy U is known since it is determined by the thermostat. We must optimize the entropy given by Eq. (1.1) with the norm constraint (1.2), and with the energy constraint

$$\sum_{i=1}^{W} p_i\, E_i = U \,. \tag{2.63}$$

Following along the lines of Section 2.3, we obtain the celebrated BG weight

$$p_i = \frac{e^{-\beta E_i}}{Z_{BG}} \,, \tag{2.64}$$

with the *partition function* given by

$$Z_{BG} \equiv \sum_{i=1}^{W} e^{-\beta E_i} \,, \tag{2.65}$$

the Lagrange parameter β being related with the temperature through $\beta \equiv 1/(kT)$.

We can prove also that

$$\frac{1}{T} = \frac{\partial S_{BG}}{\partial U} \,, \tag{2.66}$$

that the *Helmholtz free energy* is given by

$$F_{BG} \equiv U - T S_{BG} = -\frac{1}{\beta} \ln Z_{BG} \,, \tag{2.67}$$

and that the *internal energy* is given by

$$U = -\frac{\partial}{\partial \beta} \ln Z_{BG} \,. \tag{2.68}$$

In the limit $T \to \infty$ we recover the microcanonical ensemble.

2.4.3 Others

The system may be exchanging with the thermostat not only energy, so that the temperature is that of the thermostat, but also particles, so that also the chemical potential is fixed by the reservoir. This physical situation corresponds to the so-called *grand-canonical ensemble*. This and other similar physical situations can be treated along the same path, as shown by Gibbs. We shall not review here these types of systems, which are described in detail in [35], for instance.

Another important physical case, which we do not review here either, is when the particles cannot be considered as distinguishable. Such is the case of *bosons* (leading to Bose–Einstein statistics), *fermions* (leading to Fermi–Dirac statistics), and the so-called *gentilions* (leading to Gentile statistics, also called *parastatistics* [101–103], which unifies those of Bose–Einstein and Fermi–Dirac).

All these various physical systems, and even others, constitute what is currently referred to as BG statistical mechanics, essentially because at its basis we find, in one way or another, the entropic functional S_{BG}. It is this entire theoretical body that in principle we intend to generalize in the rest of the book, through the generalization of S_{BG} itself.

http://www.springer.com/978-0-387-85358-1

Introduction to Nonextensive Statistical Mechanics
Approaching a Complex World
Tsallis, C.
2009, XVIII, 382 p. 238 illus., 40 illus. in color.,
Hardcover
ISBN: 978-0-387-85358-1