Under consideration for publication in Math. Struct. in Comp. Science

Shannon entropy: a rigorous mathematical notion at the crossroads between probability, information theory, dynamical systems and statistical physics

Annick Lesne

Laboratoire de Physique Théorique de la Matière Condensée CNRS UMR 7600

Université Pierre et Marie Curie-Paris 6, 4 place Jussieu, F-75252 Paris Cedex 05, France

& Institut des Hautes Études Scientifiques

35 route de Chartres, F-91440, Bures-sur-Yvette, France.

Email for correspondence: lesne@ihes.fr

Received 18 January 2011

Statistical entropy was introduced by Shannon as a basic concept in information theory, measuring the average missing information on a random source. Extended into an entropy rate, it gives bounds in coding and compression theorems. I here present how statistical entropy and entropy rate relate to other notions of entropy, relevant either to probability theory (entropy of a discrete probability distribution measuring its unevenness), computer sciences (algorithmic complexity), the ergodic theory of dynamical systems (Kolmogorov-Sinai or metric entropy), or statistical physics (Boltzmann entropy). Their mathematical foundations and correlates (entropy concentration, Sanov, Shannon-McMillan-Breiman, Lempel-Ziv and Pesin theorems) clarify their interpretation and offer a rigorous basis to maximum entropy principles. Although often ignored, these mathematical perspectives give a central position to entropy and relative entropy in statistical laws describing generic collective behaviors. They provide insights into the notions of randomness, typicality and disorder. The relevance of entropy outside the realm of physics, for living systems and ecosystems, is yet to be demonstrated.

Contents

1. Introduction

Historically, many notions of entropy have been proposed. The etymology of the word *entropy* dates back to Clausius (Clausius 1865), in 1865, who dubbed this term from the greek *tropos*, meaning transformation, and a prefix *en-* to recall the indissociable (in his work) relation to the notion of energy (Jaynes 1980). A statistical concept of entropy was introduced by Shannon in the theory of communication and transmission of information (Shannon 1948). It is formally similar to Boltzmann entropy associated with the statistical description of the microscopic configurations of many-body systems and how it accounts for their macroscopic behavior (Honerkamp 1998; Castiglione *et al.* 2008). Establishing the relationships between statistical entropy, statistical mechanics and thermodynamic entropy was initiated by Jaynes (Jaynes 1957; Jaynes 1982b). In an initially totally different perspective, a notion of entropy rate was developped in dynamical systems theory and symbolic sequence analysis (Badii and Politi 1997; Lesne *et al.* 2009). The issue of compression is sometimes rooted in information theory and Shannon entropy, while in other instances it is rooted in algorithmic complexity (Gray 1990; Cover and Thomas 2006). As a consequence of this diversity of uses and concepts we may ask whether the use of the term entropy has any meaning. Is there really something linking this diversity, or is the use of the same term in so many meanings

just misleading? A short historical account of the different notions of entropy was given by Jaynes thirty years ago (Jaynes 1980). I here propose a more detailed overview of the relationships between the different notions of entropy, not in an historical perspective but rather as they appear today, highlighting bridges between probability, information theory, dynamical systems theory and statistical physics. I will base my argumentation on mathematical results related to Shannon entropy, relative entropy and entropy rate. They offer a reliable both qualitative and quantitative guide for the proper use and interpretation of these concepts. In particular, they provide a rationale, as well as several caveats, to the maximum entropy principle.

2. Shannon entropy

2.1. *Definitions*

For a random variable X with values in a finite set $\mathcal{X}$, *Shannon entropy* is defined as (Shannon 1948):

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) \geq 0 \tag{1}$$

It quantifies the *unevenness* of the probability distribution p. In particular, the minimum $H(X) = 0$ is reached for a constant random variable, i.e. a variable with a determined outcome, which reflects in a fully localized probability distribution $p(x_0) = 1$ and $p(x) = 0$ for $x \neq x_0$. At the opposite, $H(X)$ is maximal, equal to $\log_2(|\mathcal{X}|)$, for a uniform distribution. $H(X)$ is also denoted:

$$S(p) = -\sum_{i=1}^{|\mathcal{X}|} p(x_i)\log_2 p(x_i) \tag{2}$$

which underlines the fact that entropy is a feature of the probability distribution p. Entropy does not depend on the graph $x \to p(x)$, i.e, it is not a feature of the random variable itself but only of the set of its probability values. This property reflects in a permutation invariance of $H(X)$: let the variable $\sigma.X$ obtained by a permutation of the states, namely $\mathrm{Prob}(\sigma.X = x_{\sigma(i)}) = p(x_i)$, then $H(X) = H(\sigma.X)$. Entropy trivially increases with the number of possible states: for an unbiased coin, $H = \log_2 2 = 1$ while for an unbiased dice, $H = \log_2 6 > 1$.

According to the folklore (Avery 2003), the term entropy has been suggested to Shannon by von Neumann for both its fuzziness and resemblance with Boltzmann entropy [†]. Historically, Shannon (Shannon 1948) introduced a function $\mathcal{H}(p_1, \ldots, p_n)$ satisfying the following three requirements. Given a random variable X with values $x_1, \ldots, x_n$ and corresponding probabilities $p_1, \ldots, p_n$, with $\sum_{i=1}^n p_i = 1$:

(i) $\mathcal{H}$ is a continuous function of the p_i;

(ii) if all p_i are equal (to $1/n$), $\mathcal{H}(1/n, \ldots, 1/n)$ is a monotonous increasing function of n;

(iii) if we group $y_1 = \{x_1, \ldots, x_{k_1}\}$, $y_2 = \{x_{k_1+1}, \ldots, x_{k_1+k_2}\}$, $\ldots$, $y_m = \{x_{n-k_m+1}, \ldots, x_n\}$, so

[†] Quoting (Avery 2003): when von Neumann asked him how he was getting on with his information theory, Shannon replied that "the theory was in excellent shape, except that he needed a good name for missing information". "Why dont you call it entropy", von Neumann suggested. "In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place, no one understands entropy very well, so in any discussion you will be in a position of advantage". According to another source (Tribus and McIrvine 1971), quoting Shannon: "My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage" ".

that $q_i = \sum_{i=k_1+\ldots+k_i-1}^{i=k_1+\ldots+k_i} p_l$ is the probability of the realisation y_i, then

$$\mathcal{H}(p_1,\ldots,p_n) = \mathcal{H}(q_1,\ldots,q_m) + \sum_{i=1}^{m} q_i \mathcal{H}\left(\frac{p_{k_1+\ldots+k_{(i-1)}}}{q_i},\ldots,\frac{p_{k_1+\ldots+k_i-1}}{q_i}\right)$$

This yields a functional form $\mathcal{H}(p_1,\ldots,p_n) = -K\sum_{i=1}^{n} p_i \log_2 p_i$, which is *unique* up to the multiplicative constant K. Statistical entropy is thus almost uniquely defined according to the above natural prescriptions (i – iii). It is easy to check from the definition (2) that the entropy of a compound of independent events $Z = X_1 \ldots X_n$ such that $p_Z = p_{X_1} \ldots p_{X_n}$ is simply $H(Z) = \prod_{i=1}^{n} H(X_i)$. Another important property is *entropy convexity*. Let $p = \lambda p^0 + (1-\lambda)p^1$, then $S(p) \geq \lambda S(p^0) + (1-\lambda)S(p^1)$ or equivalently $H(X) \geq \lambda H(X^0) + (1-\lambda)H(X^1)$ where X, X^0 and X^1 are random variables with respective distributions p, p^0 and p^1. The difference $S(p) - \lambda S(p^0) - (1-\lambda)S(p^1)$ measures the uncertainty added in mixing the two distributions p_0 and p_1.

2.2. *Information-theoretic interpretation*

Shannon initially developed information theory for quantifying the information loss in transmitting a given message in a communication channel (Shannon 1948). A noticeable aspect of Shannon approach is to ignore semantics and focus on the physical and statistical constraints limiting the transmission of a message, notwithstanding its meaning. The source generating the inputs $x \in \mathcal{X}$ is characterized by the probability distribution $p(x)$. Shannon introduced the quantity $I_p(x) \equiv -\log_2 p(x)$ as a measure of the *information* brought by the observation of x knowing the probability distribution p. In plain language, one could correctly say (Balian 2004) that $I_p(x)$ is the surprise in observing x, given prior knowledge on the source summarized in p. Shannon entropy $S(p)$ thus appears as the *average missing information*, that is, the average information required to specify the outcome x when the receiver knows the distribution p. It equivalently measures the amount of uncertainty represented by a probability distribution (Jaynes 1957). In the context of communication theory, it amounts to the minimal number of bits that should be transmitted to specify x (we shall come back to this latter formulation in § 5.2 devoted to data compression and coding theorems). $I_p(x)$ is currently denoted $I(x)$, regrettably forgetting about the essential mention of the distribution p.

The actual message is one selected from a set of possible messages, and information is produced when one message is chosen from the set. A priori knowledge of the set of possible messages is essential in quantifying the information that the receiver needs in order to properly identify the message. A classical example is the quantification of the information needed to communicate a play by Shakespeare, whether the receiver knows in advance that he will receive one of the plays by Shakespeare (and then transmitting only the few first words is sufficient), or not (and then the whole text of the play has to be transmitted). What changes between the two situations is the a priori knowledge, i.e. the *set of possible messages*. In the above formalization, it is described through the a priori probability $p(x)$ describing the source. We thus emphasize that the meaning of information makes sense only with reference to the prior knowledge of the set $\mathcal{X}$ of possible events x and their probability distribution $p(x)$. It is not an absolute notion but rather a highly subjective and relative notion, and for this reason it is advisable to speak of "missing information" rather than "information". Moreover, the precise and technical meaning of information in Shannon theory often mixes up with the loose meaning of information in current language. Henceforth, we shall use the term *knowledge* instead of information when the latter is used with its non-technical meaning (plain language).

Shannon information and its statistical average, Shannon entropy, should not be confused with *Fisher information*. The latter appears in parametric estimation, that is, estimation of a parameter

a of a probability distribution $p_a(x)$. It is defined as $I_F(a) = [\partial \ln p/\partial a)]^2$ (for a one-dimensional parameter, see e.g. (Amari and Nagaoka 2000; Cover and Thomas 2006) for the multivariate extension). Its main interest comes from *Cramer-Rao bound* (Kagan *et al.* 1973), relating Fisher information and the variance $\mathrm{Var}(\hat{a})$ of the estimated value $\hat{a}$ of the parameter a according to $\mathrm{Var}(\hat{a}).I_F(a) \geq 1$. We shall discuss further the geometric meaning of Fisher information in relation to relative entropy in § 2.4.

2.3. *Conditional entropy, relative entropy and Kullback-Leibler divergence*

Shannon entropy extends to multivariate random variables (Gray 1990; Cover and Thomas 2006). It involves their joint distribution, e.g. for two random variables X and Y taking their values in a priori two different (discrete and finite) spaces $\mathcal{X}$ and $\mathcal{Y}$:

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y) \tag{3}$$

From this joint entropy, one defines the *conditional entropy* $H(X\,|\,Y) \equiv H(X,Y) - H(Y)$, which appears to be the average (over Y) of the entropies of the conditional probability distributions $p(X\,|\,Y=y)$:

$$H(X\,|\,Y) \equiv H(X,Y) - H(Y) = \sum_{y\in\mathcal{Y}} p(y)\ \left[-\sum_{x\in\mathcal{X}} p(x|y)\log_2 p(x|y)\right] \tag{4}$$

Using a convexity argument (Honerkamp 1998), it is straightforward to show that $H(X\,|\,Y) \leq H(X) \leq H(X,Y)$. In particular, $H(X|Y) \leq H(X)$ reflects the fact that uncertainty on X is never increased by the knowledge of Y. In case of multiple conditioning, we have $H(X|Y,Z) \leq H(X|Y) \leq H(X)$.

When the random variables X and Y have the same state space $\mathcal{X}$, with respective distributions p_X and p_Y, it is possible to consider the *relative entropy*:

$$H_{rel}(X|Y) \equiv S_{rel}(p_X|p_Y) = -\sum_x p_X(x)\log_2[p_X(x)/p_Y(x)] \tag{5}$$

It is easy to show (Cover and Thomas 2006) that $S_{rel}(p_X|p_Y) \leq 0$. The opposite of the relative entropy defines the *Kullback-Leibler divergence* (Kullback and Leibler 1951). For two probability distributions p and q on the same space $\mathcal{X}$:

$$D(p||q) = -S_{rel}(p|q) = \sum_x p(x)\log_2[p(x)/q(x)] \geq 0 \tag{6}$$

$D(p||q)$ is not a distance since it is not symmetric and does not satisfy the triangular inequality; its sole common property with a distance is that $D(p||q) \geq 0$ and $D(p||q) = 0$ if and only if $p = q$. We shall see (§ 2.4 and § 3.2) that it has nevertheless useful geometric properties and interpretation.

To give an illustration of the use and interpretation of these quantities, let us consider a stationary Markov chain $(X_t)_{t\geq 0}$. Then $H(X_t|X_0)$ and $H(X_0|X_t)$ increase with time t, while $D(p_t||p_{stat})$ decreases to 0 (denoting p_t the distribution at time t and p_{stat} the stationary distribution). We emphasize that one should not confuse:

— the conditional entropy $H(X|Y) = H(X,Y) - H(Y)$ of the random variables X and Y, which could now take their values in different sets $\mathcal{X}$ and $\mathcal{Y}$; its computation requires to know the joint distribution p_{XY} of X and Y, defined on the product space $\mathcal{X}\times\mathcal{Y}$.

— the relative entropy $H_{rel}(X|Y)$ between the random variables X and Y, taking their values in the *same set* $\mathcal{X}$, or equivalently Kullback-Leibler divergence $D(p_X||p_Y)$ between their probability distributions both defined on $\mathcal{X}$;

The distinction between relative and conditional entropies becomes yet clearer when introducing the *mutual information*:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{7}$$

of two random variables X and Y. Its definition could be reformulated as:

$$I(X;Y) = D(p_{XY}||p_X p_Y) \tag{8}$$

where p_{XY} is the joint distribution of (X,Y) and p_X, p_Y the marginal distributions. Mutual information $I(X;Y)$ measures the full correlations between X and Y: it vanishes if and only if X and Y are independent, while it equals $H(X)$ if $X = Y$. The notion can be extended into a *conditional mutual information* (mutual information between X and Y given Z), defined as:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{9}$$

2.4. *Information geometry*

Kullback-Leibler divergence has another geometric meaning, related to the so-called *Fisher information metric* (Amari and Nagaoka 2000). For a parametrized family $p(x,a)$ of probability distributions where a has d components, Fisher information metric writes:

$$dp^2(a) = \sum_{i,j} g_{i,j}(a) da^i da^j \quad \text{where} \quad g_{ij}(a) = \int \frac{\partial \log_2 p(x,a)}{\partial a_i} \, \frac{\partial \log_2 p(x,a)}{\partial a_j} \, p(x,a) dx \tag{10}$$

hence

$$dp^2(a) = 2D[p(.,a)||p(.,a+da)] \tag{11}$$

This metric endows the parametrized family with a d-dimensional Riemannian differential manifold structure. The parameters a give the coordinates on this manifold. Working at the level of a space of probability distributions recovers a continuous setting even if the underlying features (e.g. state space) are discrete. This makes available the tools of differential geometry, allowing to study statistical structures as geometrical structures.

Parameters a are currently to be estimated from data. As already mentioned above for $d = 1$, a bound on the estimator is given by the *Cramer-Rao theorem*, stating that (Kagan *et al.* 1973): $V_a(\hat{a}) - G(a)^{-1}$ is a positive semi-definite matrix, where $V_a(\hat{a})$ is the variance-covariance matrix of the estimator $\hat{a}$ with respect to the distribution $p(.,a)$ and $G(a)$ is the Fisher metric at point $p(.,a)$. It means that the local geometry of the space of probability distributions, as described by the Fisher information metric, expresses the sensitivity of the distributions with respect to their parameters, which is relevant both for estimation from experimental data and control.

2.5. *Behavior of entropy upon coarse-graining and local averaging*

Compositional law requirement (iii) involved in the construction of Shannon entropy, § 2.1, expresses explicitly the behavior of Shannon entropy upon a *coarse-graining*, in which elementary states $(x_i)_{i=1,\dots,n}$ are grouped into macrostates $(y_j)_{j=1,\dots,m}$, with $\text{Prob}(y) = \sum_{x \in y} p(x)$. The entropy $H(X)$ of the original variable X is equal to the entropy of the coarse-grained variable Y, supplemented with the average entropy of a grain $Y = y_j$, that is, the conditional entropy of X knowing $Y = y_j$ averaged over the probability distribution of Y. This additional term is nothing but the conditional entropy $H(X \,|\, Y)$), from which follows that:

$$H(X) = H_{cg}(Y) + H(X \,|\, Y) \tag{12}$$

$H_{cg}(Y) \leq H(X)$, with a strict inequality as soon as the coarse-graining is non trivial, i.e. $m < n$.

On the other hand, Shannon in its seminal paper (Shannon 1948) already noticed that any change toward equalization of the probabilities $p_1, \ldots, p_n$ increases H. Such a change is in particular achieved through a *local averaging*. For instance, defining $p'_i = [p_i + p_{i+1}]/2$ (with $p'_n = (p_n + p_1)/2$), one gets $H_{av} = S(p') \geq H = S(p)$. Local averaging should not be confused with coarse-graining. Local averaging preserves the number of elements and $H_{av}(\epsilon)$ increases with the scale ϵ at which the local averaging is performed ($\epsilon = 2$ in the above example). By contrast, coarse-graining amounts to grouping elements into a reduced number of macro-elements or grains. It leads to an entropy decrease : $H_{cg} \leq H_0$, where the decrease gets stronger when the size ϵ of the grains increases.

Another puzzling situation is the case where the transmission in a communication channel is incomplete, and yields X as the outcome of an input (X, Y). The first way to model this situation is to describe a deterministic channel truncating the initial message. In this viewpoint, the entropy of the output is $H(X)$, which is lower than the entropy $H(X, Y)$ of the source, and incomplete transmission would be said to decrease entropy. A second way to model the situation is to describe a noisy channel replacing Y by a fully random noise η with the same state space $\mathcal{Y}$ and fully independent of X. Now the entropy of the output is $H(X)+H(\eta) = H(X)+\log_2 |\mathcal{Y}|$ which is larger than $H(X)+H(Y)$, itself larger than the entropy $H(X, Y)$ of the source: the truncated transmission corresponds to an increase of entropy. Entropy is thus dramatically sensitive to the considered set of possible events, here $\mathcal{X}$ or $\mathcal{X} \times \mathcal{Y}$ respectively. This example underlines the irrelevance to speak of an information loss or information transfer between systems having different states spaces. We here recover the caveat that speaking of (missing) information makes sense only with respect to the prior knowledge of the possible states.

2.6. *Continuous extension*

Extension of entropy to continuous-valued random variables has already been discussed in (Shannon 1948) and it is today a textbook matters (Ihara 1993). The entropy now expresses $S(p) = -\int_{\mathcal{X}} p(x) \log_2 p(x) dx$ where $p(x)$ is a density with respect to the measure dx. The difficulty in extending entropy to a random variable taking its values in a continuous set comes from the fact that$-\int dx\, p(x) \log_2 p(x)$ is not invariant upon a change of coordinate $y = f(x)$, leading to replace $p(x)dx$ by $q(y)dy$ with $p(x) = |f'(x)|q(f(x)))$. Whereas the discrete entropy is an absolute quantity, this continuous entropy is relative to a coordinate system, and defined up to an additive constant. The difficulty disappears when considering the relative entropy or equivalently the Kullback-Leibler divergence (Ihara 1993) since the continuous extension $D(p||q) = \int_{\mathcal{X}} p(x) \log_2[p(x)/q(x)]dx$ is now invariant upon a change of coordinates [‡]. We here see an instance of the general fact that continuous probabilities are fundamentally more delicate to tackle, leading to well-known paradoxes, like the Bertrand paradox (namely, what is the probability that a long needle drawn at random intersects a given circle with a chord longer than a given length). The main point is to keep in mind that the meaningful quantity, having a proper mathematical behavior e.g. under a change of coordinates, is not $p(x)$ but $p(x)dx$.

3. Concentration theorems and maximum entropy principles

3.1. *Types and entropy concentration theorems*

We here consider sequences $\bar{x}_N \in \mathcal{X}^N$ of N independent and identically distributed random variables with values in a finite set $\mathcal{X}$. This space $\mathcal{X}$ is currently termed the *alphabet* and its elements

[‡] Actually $D(\mu||\mu_0)$ is defined for any pair of probability measures on a Polish topological space $\mathcal{X}$, e.g. a closed subset of $\mathbf{R}^d$, provided the probability measure μ is absolutely continuous with respect to μ_0. Then it writes $D(\mu||\mu_0) = \int_{\mathcal{X}} d\mu \log(d\mu/d\mu_0)$; otherwise $D(\mu||\mu_0) = +\infty$.

symbols, in reference to messages in communication theory. The definitions and results of this section will equally apply to configurations of N independent and identical elements with elementary states in $\mathcal{X}$. A first essential notion is the *type* $p_{\bar{x}_N}$ of the sequence or configuration $\bar{x}_N$: it is the relative number of occurrences of each symbol in $\bar{x}_N$. In other words, it is the *empirical distribution* of the symbols in the sequence $\bar{x}_N$. It is thus an observable quantity, derived in an observed sequence $\bar{x}_N$ as the *normalized histogram* of the different symbols.

The sequence space $\mathcal{X}^N$ can be partitioned into classes of sequences having the same type. By extension, these classes are termed "types". Each probability distribution p on $\mathcal{X}$ defines a type, i.e. a subset of $\mathcal{X}^N$. The space $\mathcal{X}^N$ can be seen as a microscopic phase space (see § 8.2); the types then define macrostates. There are at most $(1+N)^{|\mathcal{X}|}$ different types (Cover and Thomas 2006), whereas the number of sequences in $\mathcal{X}^N$ grows exponentially. In consequence, at least one type has exponentially many elements (Csiszár 1998). Actually, we shall see now that, asymptotically, one type contains most elements. In its simplest formulation, the *entropy concentration theorem* states (Georgii 2003) (Card denotes the cardinal):

$$\lim_{N\to\infty} \frac{1}{N} \log_2 \mathrm{Card}\{\bar{x}_N \in \mathcal{X}^N, p_{\bar{x}_N} = p\} = H(p) \tag{13}$$

extending into the relaxed statement, for any sequence p_N tending to p as $N \to \infty$:

$$\lim_{N\to\infty} \frac{1}{N} \log_2 \mathrm{Card}\{\bar{x}_N \in \mathcal{X}^N, p_{\bar{x}_N} = p_N\} = H(p) \tag{14}$$

Denoting Prob the equiprobable distribution on the microscopic phase space $\mathcal{X}^N$ (i.e. the normalized cardinal), this statement can be rewritten:

$$\lim_{N\to\infty} \frac{1}{N} \log_2 \mathrm{Prob}[\bar{x}_N \in \mathcal{X}^N, p_{\bar{x}_N} = p_N] = H(p) - \log_2 |\mathcal{X}| \tag{15}$$

It accommodates some fixed tolerance ϵ, namely:

$$\lim_{N\to\infty} \frac{1}{N} \log_2 \mathrm{Prob}[\bar{x}_N \in \mathcal{X}^N, |p_{\bar{x}_N} - p| < \epsilon] = H(p) - \log_2 |\mathcal{X}| \tag{16}$$

Let p^* the distribution maximizing Shannon entropy: $H(p^*) = \log_2 |\mathcal{X}|$ whereas $H(p) - \log_2 |\mathcal{X}| < 0$ for any other type $p \neq p^*$. It means that asymptotically, the type of p^* contains almost all sequences. More precisely, one can show:

$$\lim_{N\to\infty} \mathrm{Prob}[\bar{x}_N \in \mathcal{X}^N, |p_{\bar{x}_N}(x) - p^*| < \epsilon] = 1 \tag{17}$$

Configurations with type p^* form a *typical set*: it is exponentially large compared to any other set containing sequences with type p with $p \neq p^*$:

$$\lim_{N\to\infty} \mathrm{Prob}[\bar{x}_N \in \mathcal{X}^N, p_{\bar{x}_N}(x) = p] = 0 \quad \text{and} \quad \lim_{N\to\infty} \mathrm{Prob}[\bar{x}_N \in \mathcal{X}^N, |p_{\bar{x}_N}(x) - p^*| > \epsilon] = 0 \tag{18}$$

decreasing exponentially fast with N as stated above in (15). The type of p^*, although typical, is nevertheless exponentially small compared to the set of all possible configurations, which underlies the difference between *possible* and *probable* configurations.

These statements extend to the case of a constrained subset $\mathcal{D}$ of probability distributions:

$$\mathcal{D} = \{p \text{ probability on } \mathcal{X}, \langle a_\alpha(X)\rangle_p = A_\alpha, \quad \alpha = 1, \ldots, m \} \tag{19}$$

The statistical average $\langle a_i(X)\rangle$ computed with respect to the empirical distribution $p_{\bar{x}_N}$ is nothing but the empirical average $(1/N)\sum_{i=1}^N a_\alpha(x_i)$. It is thus straightforward to check whether a given observation $\bar{x}_N$ (actually a set of independent individual observations) satisfy the constraints, i.e. whether its type belongs to $\mathcal{D}$. Denoting $p^*_{\mathcal{D}}$ the distribution maximizing Shannon entropy in $\mathcal{D}$, the conditional probability in $\mathcal{X}^N$ that the type of a sequence is close to $p^*_{\mathcal{D}}$, within some fixed

tolerance $\epsilon > 0$, converges to 1:

$$\lim_{N\to\infty} \text{Prob}[\bar{x}_N \in \mathcal{X}^N, |p_{\bar{x}_N}(x) - p^*_{\mathcal{D}}(x)| < \epsilon \mid p_{\bar{x}_N} \in \mathcal{D}] = 1 \tag{20}$$

Entropy concentration theorem describes quantitatively the fact that almost all configurations satisfying empirically the constraints (that is, having empirical averages equal to A_α, $\alpha = 1, \ldots, m$) have an empirical distribution *asymptotically close to the maximum entropy distribution* $p^*_{\mathcal{D}}$. The same statement holds with relaxed constraints, defining a larger set of probability distributions:

$$\mathcal{D}_\delta = \{p \text{ probability on } \mathcal{X}, |\langle a_\alpha(X)\rangle_p - A_\alpha| < \delta, \quad \alpha = 1, \ldots, m \} \tag{21}$$

and replacing $p^*_{\mathcal{D}}$ with the distribution $p^*_{\mathcal{D}_\delta}$ maximizing Shannon entropy in $\mathcal{D}_\delta$. One can show (Robert 1990) that $p^*_{\mathcal{D}}$ and $p^*_{\mathcal{D}_\delta}$ are unique, and that $p^*_{\mathcal{D}_\delta}$ weakly converges to $p^*_{\mathcal{D}}$ when δ converges to 0.

Since uncorrelated sequences are not always a realistic model, one may wonder about a concentration theorem for correlated sequences. We shall see such an extension below, § 5.1. The main modification required to capture correlations is to replace H by an average entropy rate h.

3.2. *Relative entropy concentration and Sanov theorems*

We could also be willing to make statement about the *asymptotic weight* of the different types, given the distribution p_0 of the elements. Accordingly, we now consider sequences of independent elements whose states are identically distributed according to the distribution p_0 on $\mathcal{X}$ (sometimes termed the *reference distribution*). The sequences in $\mathcal{X}^N$ being uncorrelated, their probability distribution is merely the product distribution $p_0^{\otimes N}$. In this case, all the sequences with the same type have the same probability since (Georgii 2003):

$$p_0^{\otimes N}(\bar{x}_N) = 2^{-N[H(p_{\bar{x}_N}) + D(p_{\bar{x}_N}||p_0)]} \tag{22}$$

This identity shows that the quantity controlling the asymptotic behavior is no longer the entropy but the relative entropy, or equivalently its opposite, the Kullback-Leibler divergence (6):

$$\lim_{N\to\infty} \frac{1}{N} \log_2 p_0^{\otimes N}[\bar{x}_N \in \mathcal{X}^N, p_{\bar{x}_N} = p] = -D(p||p_0) \tag{23}$$

recovering (15) when p_0 is uniform; indeed, if p_{unif} is the equiprobable distribution on $\mathcal{X}$, then $D(p||p_{unif}) = \log_2 |\mathcal{X}| - H(p)$. This statement accommodate some fixed tolerance $\epsilon > 0$, namely:

$$\lim_{N\to\infty} \frac{1}{N} \log_2 p_0^{\otimes N}[\bar{x}_N \in \mathcal{X}^N, |p_{\bar{x}_N} - p| < \epsilon] = -D(p||p_0) \tag{24}$$

It embeds the well-known estimation theorem stating that the empirical distribution $p_{\bar{x}_N}$ converges to the actual one p_0. In particular, the law of large numbers ensures that almost surely $\lim_{N\to\infty} D(p_{\bar{x}_N}||p_0) = 0$ (Csiszár and Körner 1981; Csiszár 1998). But the above statements go further and allow to control the remainder, that is, large deviations, finite-size errors in the estimation and the convergence rate towards the true distribution p_0.

A related issue is the inference of distributions satisfying linear constraints, typically the knowledge of some moments, while imposing the least possible bias on the reference distribution p_0. Considering the same constrained sets $\mathcal{D}_\delta$ and $\mathcal{D}$ as in § 3.1, the solution is given by the closest distributions to p_0, as measured by the Kullback-Leibler divergence, that belong respectively to $\mathcal{D}_\delta$ and $\mathcal{D}$. More precisely, under the assumption that $\mathcal{D}$ is not empty and contains at least one distribution having a non-vanishing relative entropy with respect to p_0, one can prove that (Robert 1990):

(i) there exists a unique distribution $p^*_{\mathcal{D}_\delta}$ in $\mathcal{D}_\delta$ and a unique distribution $p^*_{\mathcal{D}}$ in $\mathcal{D}$ maximizing the relative entropy with respect to the reference distribution p_0, respectively in $\mathcal{D}_\delta$ and $\mathcal{D}$;

(ii) $p^*_{\mathcal{D}_\delta}$ weakly converges to $p^*_{\mathcal{D}}$ when δ converges to 0;

(iii) $p^*_{\mathcal{D}_\delta}$ has the concentration property in $\mathcal{D}_\delta$: for any neighborhood $\mathcal{V}_\delta$ of $p^*_{\mathcal{D}_\delta}$ in $\mathcal{D}_\delta$ (for the narrow topology, i.e. the weak topology for bounded continuous real functions on $\mathcal{X}$), $\exists \alpha > 0$, $\exists N_0$, such that $\forall N \geq N_0$, $p_0^{\otimes N}[\bar{x} \in \mathcal{X}^N, p_{\bar{x}_N} \notin \mathcal{V}_\delta \mid p_{\bar{x}_N} \in \mathcal{D}_\delta] \leq e^{-N\alpha}$.

(iv) $p^*_{\mathcal{D}}$ has the concentration property in $\mathcal{D}$: for any neighborhood $\mathcal{V}$ of $p^*_{\mathcal{D}}$ in $\mathcal{D}$, $\exists \alpha > 0$, $\exists N_0$, such that $\forall N \geq N_0$, $p_0^{\otimes N}[\bar{x} \in \mathcal{X}^N, p_{\bar{x}_N} \notin \mathcal{V} \mid p_{\bar{x}_N} \in \mathcal{D}] \leq e^{-N\alpha}$.

This is a *large deviation* result (Ellis 1985; Touchette 2009), stated in a way supplementing the previous concentration theorem, § 3.1, since it allows to control the remainder. It is valid only with Kullback-Leibler divergence: other measures of the distance between the two distributions, e.g. quadratic distance, do not satisfy a large deviation statement (Robert 1990). Note that $D(p^*_{\mathcal{D}_\delta}||p_0) \leq D(p^*_{\mathcal{D}}||p_0)$ since $\mathcal{D} \subset \mathcal{D}_\delta$. Let us consider the distribution σ^* maximizing the relative entropy $-D(\sigma||p_0)$ in the complement of $\mathcal{V}$ in $\mathcal{D}$ (thus by construction $D(\sigma^*||p_0) > D(p^*_{\mathcal{D}}||p_0)$). The exponent α is roughly given by $D(\sigma^*||p_0) - D(p^*_{\mathcal{D}}||p_0)$. The larger is $\mathcal{V}$, the more distant σ^* is from $p^*_{\mathcal{D}}$, the larger is $D(\sigma^*||p_0)$ hence the larger is its difference with $D(p^*_{\mathcal{D}}||p_0)$, and the larger is α. It means that the exponential decrease as N tends to infinity of the relative weight of the configurations whose type is not in $\mathcal{V}$ is the faster the larger is $\mathcal{V}$, i.e. the more distant these types are from $p^*_{\mathcal{D}}$. When $\mathcal{X}$ is discrete and p_0 is uniform (all states in $\mathcal{X}$ having the same probability $1/|\mathcal{X}|$), the probability $p_0^{\otimes}$ coincides with the normalized cardinal and we recover the particular theorem derived by Jaynes (Jaynes 1982a).

For independent random variables identically distributed according to the distribution p_0, the above statements extend to convex subsets $\mathcal{C}$ of probability distributions on $\mathcal{X}$, according to the *Sanov theorem* (Sanov 1957):

$$\lim_{N\to\infty} \frac{1}{N} \log_2 p_0^N(\bar{x}_N \in \mathcal{X}^N, p_{\bar{x}_N} \in \mathcal{C}) = -\inf_{\nu\in\mathcal{C}} D(\nu\,||, p_0) \tag{25}$$

It is a large deviation result (Ellis 1985; Touchette 2009), that could also be seen as a projection (Georgii 2003), see below § 3.3. Sanov theorem extends to continuous densities. The space $\mathcal{X}$ is now a continuous metric space. For any convex subset $\mathcal{C}$ of probability densities on $\mathcal{X}$:

$$\lim_{N\to\infty} (1/N) \log_2 g^{\otimes N}[\bar{x}_N \in \mathcal{X}^N, \phi_N(x) \in \mathcal{C}) = -\inf_{f\in\mathcal{C}} D(f\,||\,g) \tag{26}$$

where $\phi_N(x) = (1/N)\sum_{i=1}^N \delta(x - x_i)$ is the empirical distribution (continuous type) and for instance $\mathcal{C}$ is defined according to

$$\phi(x) \in \mathcal{C} \Longleftrightarrow \int_{\mathcal{X}} h(x)\phi(x)dx = \alpha \tag{27}$$

The theorem states that the major contribution to $g^{\otimes N}[\bar{x}_N \in \mathcal{X}^N, \phi_N(x) \in \mathcal{C}]$ comes from the distribution that maximizes the relative entropy under the constraint of belonging to $\mathcal{C}$.

3.3. *Geometric interpretation*

Kullback-Leibler divergence is a useful tool allowing to work in a space of probability distributions, here the space $\mathcal{P}(\mathcal{X})$ of probability distributions on $\mathcal{X}$. For instance, for any subset $\mathcal{C} \subset \mathcal{P}(\mathcal{X})$ that does not contain the reference distribution p_0, we could consider:

$$Q(p_0) = \text{argmin}_{q\in\mathcal{P}(\mathcal{X})} D(q||p_0) \tag{28}$$

The distributions in $\mathcal{C}$ minimizing $D(.||p^0)$ could be termed the "orthogonal projections" of p^0 onto $\mathcal{C}$. When $\mathcal{C}$ is closed for the weak topology, such minimizers are ensured to exist (Georgii 2003). If moreover $\mathcal{C}$ is convex, then the minimizer $Q(p_0)$ is uniquely determined (Csiszár 1975). It is called the *I-projection* of p^0 on $\mathcal{C}$ and $D(Q(p_0)||p^0)$ measures the "distance" between p^0 and the set $\mathcal{C}$.

For this reason, Kullback-Leibler divergence could be more natural and more efficient than true functional distances, e.g. L_2 distance $||p-q||_2$. Recall that the mapping $p \to D(p||p^0)$ is lower semi-continuous in this weak topology (Georgii 2003); it is also strictly convex.

Given a sample of probability distributions $(p_k)_{k=1,\dots,m}$, Kullback-Leibler divergence allows to define a notion of *empirical average* $\bar{p}$ of the sample, according to (Georgii 2003; Balding *et al.* 2008)

$$\bar{p} = \text{argmin}_{q \in \mathcal{P}(\mathcal{X})} \frac{1}{m} \sum_{k=1}^{m} D(q||p_k) \tag{29}$$

This definition of an average object is suitable when arithmetic mean of the sample does not make any sense (Balding *et al.* 2008); it shows that the average can in fact be identified with a projection.

Another interesting interpretation of Kullback-Leibler divergence is related to *conditional expectation*. Let Σ be a sub-sigma algebra of the original one Σ_0. Let p be the original probability density of the considered random variable X, defined on Σ_0. The conditional expectation $E^\Sigma p$ is the Σ-measurable density such that for any Σ-measurable function f:

$$\int f(x)p(x)dx = \int f(x)E^\Sigma p(x)dx \tag{30}$$

$E^\Sigma p$ corresponds to the coarse-graining of p adapted to the coarser sigma-algebra Σ, i.e. a coarse description of the random variable X. An explicit computation, using that $\int p(x)\log_2[E^\Sigma p(x)]dx = \int E^\Sigma p(x)\log_2[E^\Sigma p(x)]dx$ according to the above definition, leads straightforwardly to the relation:

$$D(p||E^\Sigma p) = S(E^\Sigma p) - S(p) \geq 0 \tag{31}$$

Moreover, it comes $D(p||q) - D(p||E^\Sigma p) = D(E^\Sigma p||q)$ for any Σ-measurable density q, hence (Csiszár 1975):

$$\text{argmin}_{q\ \Sigma-\text{measurable}}\, D(p||q) = E^\Sigma p \tag{32}$$

3.4. *Maximum-entropy inference of a distribution*

An acknowledged heuristic principle in constructing a statistical model given some prior knowledge, e.g. experimental data, is to minimize the bias introduced in the reconstruction: an observer with the same (in general partial) knowledge would make the same inference (Bricmont 1995). In particular, without any prior knowledge on the observed process, one should consider equiprobable outcomes. This principle dates back to *Laplace principle of indifference* (or *principle of insufficient reason*) (Jaynes 1979; Jaynes 1982a; Jaynes 1982b). When the constraints are linear with respect to the probability distribution, e.g. a condition on its support and/or prescribed values for some of its moments, a constructive method for implementing this principle is to maximize Shannon entropy under the constraints. Reconstruction of the probability distribution using maximum entropy principle is by no means restricted to statistical mechanics or any other specific applications: it is a general method of inference under a priori constraints, ensuring that no additional arbitrary assumptions, i.e. no bias, are introduced (Frank 2009). A discrepancy between predictions and observations presumably originates from an ill-constrained maximum entropy principle, and gives evidence for the need of additional constraints (or the need of relaxing spurious constraints). Constraints amount to restrict the relevant space of probability distributions in which the statistical model is to be reconstructed. Once the constrained probability space is given, the distribution achieving maximum entropy is unique, due to the concavity of the entropy. Indeed, if p_1^* and p_2^* were two distinct distributions achieving the maximum entropy value H^*, any convex combination $\lambda p_1^* + (1-\lambda)p_2^*$ with $0 < \lambda < 1$ would achieve a strictly larger value $H(\lambda p_1^* + (1-\lambda)p_2^*) > \lambda H(p_1^*) + (1-\lambda)H(p_2^*) = H^*$.

More explicitly, let us consider a random variable X having n possible outcomes $x_1, \ldots, x_n$. We do not know the corresponding probabilities $p(x_1), \ldots, p(x_n)$ but only the value of some averages $\langle a_\alpha(X)\rangle_p = \sum_{i=1}^n p(x_i)a_\alpha(x_i)$, $\alpha = 1, \ldots, m$, and we want to estimate another average $\langle b(X)\rangle$ (which is exactly exactly the issue encountered in statistical mechanics, § 8). As just explained, the problem can be reformulated as follows: what is the distribution $p(x)$ on the finite space $\mathcal{X} = \{x_1, \ldots, x_n\}$ maximizing Shannon entropy under the constraints:

(i) $p(x) \geq 0$ for any $x \in \mathcal{X}$,

(ii) $\sum_{x\in\mathcal{X}} p(x)dx = 1$,

(iii) for $\alpha = 1, \ldots, m$, $\sum_{x\in\mathcal{X}} p(x)a_\alpha(x)dx = A_\alpha$.

The solution writes (Jaynes 1982a; Frank 2009):

$$p(x_j) = C\ \exp\left(-\sum_{\alpha=1}^{m} \lambda_\alpha a_\alpha(x_j)\right) \tag{33}$$

where the Lagrange multipliers λ_α are determined so that the constraints (iii) are satisfied and the multiplicative constant C ensures the proper normalization (ii). This solution has already been established by Shannon (Shannon 1948) for continuous distributions in some specific contexts: the distribution on $[-\infty, \infty[$ maximizing entropy at fixed average μ and fixed variance σ^2 is the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, and the distribution on $[0, \infty[$ maximizing entropy at fixed average μ is the exponential distribution $f(x) = (1/\mu)e^{-x/\mu}$.

Underlying maximum entropy principle is the view of probabilities as an expression of our ignorance: maximum entropy approach belongs to the subjective probability viewpoint as opposed to the frequentist viewpoint (Gillies 2000). As discussed below in § 9.1, probability distributions represent our state of knowledge, rather than an intrinsic feature or behavior of the observed system. There should be no confusion between maximum entropy principle and maximum entropy *production* principle (if any). Entropy production represents the part of the thermodynamic entroppy variation due to irreversible process, letting apart the contribution of matter transfer (§ 8.5), and it has no direct statistical counterpart.

Prior knowledge on the system in the absence of constraints expresses under the form of a reference probability distribution p_0, reflecting for instance symmetries and invariance properties. This additional knowledge has to be taken into account in statistical inference of the actual probability distribution by maximizing the relative entropy $S_{rel}(p|p_0)$ under constraints, or equivalently minimizing the Kullback-Leibler divergence $D(p||p_0) = -S_{rel}(p|p_0)$ under constraints (Banavar *et al.*. 2010). This turns the maximum entropy principle into a *maximum relative entropy principle*. Note that while maximum entropy principle applies to discrete distributions, maximum relative entropy principle also applies to continuous distributions with no ambiguity (§ 2.6). The density maximizing relative entropy under the constraint $\langle h(X)\rangle = \alpha$ is given by (up to a normalization factor) (Kagan *et al.* 1973):

$$f_\lambda(x) \sim g(x)e^{\lambda h(x)} \tag{34}$$

and λ is such that:

$$\int_{\mathcal{X}} h(x)f_\lambda(x)dx = \alpha \tag{35}$$

A similar statement holds with array of conditions, in which case $h(x)$ and λ are vectors, and the product has to be replaced by a scalar product $\lambda^t.h(x)$. A justification of the maximum relative entropy principle has been given by Van Campenhout and Cover (Van Campenhout and Cover 1981): the distribution maximizing the relative entropy can be characterized as the limit of a sequence of conditional probabilities. Let us consider independent and identically distributed

variables $X_1, \ldots, X_N$ with density p_0. The conditional probability of X_1 under the constraint $(1/N)\sum_{i=1}^{N} h(X_i) = \alpha$ on the empirical average converges to the distribution maximizing the relative entropy, namely:

$$\lim_{N\to\infty} d\mathrm{Prob}\left(X_1 = x \,|\, \frac{1}{N}\sum_{i=1}^{N} h(X_i) = \alpha \right) = f_\lambda(x)dx \tag{36}$$

The usefulness of this theorem depends largely on the proper identification of the observable $h(x)$. The ideal situations is to identify a function $h(x)$ such that $\bar{h}_N = (1/N)\sum_{i=1}^{N} h(X_i)$ is a sufficient statistics, summarizing all the relevant knowledge about the sample at the macroscopic scale. In this case, given $\bar{h}_N$, the sample associated with the maximum relative entropy distribution is maximally random and conveys no further information.

Note the *transitivity* of the maximum relative entropy principe. Namely, starting from a reference distribution p_0, we could first determine the maximum relative entropy distribution p_1^* associated with a set of constraints $\langle\varphi_i\rangle = a_i$, $i = 1, \ldots, n$, then starting with p_1^* as the reference distribution, we could determine the maximum relative entropy distribution p_2^* associated with a set of constraints $\langle\varphi_i\rangle = a_i$, $i = n+1, \ldots, n+m$. It would be the same as determining the maximum relative entropy distribution associated with the set of constraints $\langle\varphi_i\rangle = a_i$, $i = 1, \ldots, n+m$ starting from the reference distribution p_0.

A serious conceptual problem in the practical application of maximum entropy inference method has been pointed out in (Haegeman and Etienne 2010): the distribution obtained by maximizing entropy seems to depend on the chosen setting. Let us consider for instance a system composed of M cells and N individuals, and investigate the distribution of the pattern formed by the partition of individuals in the different cells. In a first viewpoint, the system configuration is described by labelling the individuals and the cells and recording the cell m_i in which the individual i lies; we thus obtain a N^M possible configurations $\mathbf{m} = (m_1, \ldots, m_N)$. In a second viewpoint, the M cells are labelled but individuals are now indiscernible, and the system configuration is described by the occupancy numbers n_m of the cells; we thus obtain at most $M^N \ll N^M$) configurations $\mathbf{n} = (n_1, \ldots, n_M)$. Note that the later description follows from a coarse-graining of the former, according to $n_m = \sum_{i=1}^{N} \delta(m, m_i)$. Maximum entropy principle applied to inference of the distribution $p(\mathbf{m})$ yields a uniform distribution, for which all configurations $\mathbf{m}$ are equiprobable, and the maximal entropy is $S = N \log_2 M$. By contrast, maximum entropy principle applied to inference of the distribution $p(\mathbf{n})$ yields a uniform distribution for the coarse-grained configuration $\mathbf{n}$, which obviously does not coincide with the coarse-grained distribution obtained from an equipartition of the elementary configurations $\mathbf{m}$. This discrepancy is quite puzzling. It means that accounting or not from the identity of the individuals is an information that strongly modifies the inference; another clue about this difference comes from the different levels of description and the fact that entropy does not commute with coarse-graining (the entropy of a coarse-grained description being always smaller than the entropy computed at a more refined level). This leads to the open question of the proper a priori choices in using maximum entropy inference method, since the choice of the configuration space strongly influence the resulting distribution and the macroscopic quantities (averages and moments) that could be computed from it. Paradoxes are solved by a consistency argument (constraints consistent with the considered distribution) or turning back to the mathematical foundations (types and entropy concentration theorems). Here, concentration theorems apply only to the convergence of the empirical distribution $\mathbf{n}$ of the population among the different spatial cells towards the actual spatial distribution (spatial type). By contrast, there is no rigorous mathematical foundation for the application of maximum entropy principle neither to the reconstruction of the distribution $p(\mathbf{m})$ nor that of $p(\mathbf{n})$.

We emphasize that maximum entropy principle is justified not only as the formalization of the in-

tuitive indifference principle of Laplace but also, rigorously, by the entropy concentration theorems, § 3.1. These theorems state that asymptotically, i.e. for a large enough number N of independent and identical elements with elementary states $x \in \mathcal{X}$, the number of configurations whose empirical distribution (their type) is the probability distribution p on $\mathcal{X}$ behaves as $2^{NH(p)}$. In consequence, an exponentially dominant set of configurations yields the distribution p^* achieving the maximum entropy. Observing experimentally a microscopic configuration yielding a different empirical distribution $p_{obs} \neq p^*$ has an exponentially small probability, decreasing like $2^{-[H(p^*)-H(p_{obs}]}$. The statement extends to constrained distributions: the configurations whose type satisfies a set of linear constraints concentrate about the distribution maximizing relative entropy under the same constraints. Accordingly, almost all microscopic configurations behave in the same way as regards their macroscopic features. It is thus legitimate to predict that real distributions in $\mathcal{X}$ will almost surely agree with the prediction of the maximum entropy principle, in the limit $N \to \infty$. A similar statement holds when a reference distribution p_0 is given, replacing Shannon entropy $H(p)$ with the relative entropy $-D(p||p_0)$. The asymptotic nature of the statement, reflecting in a condition on the sample size N, could nevertheless be limiting in some situations. Also, it relies on the independence of the different elements or individuals which often cannot be assumed. We shall see below, § 5.1, an extension of concentration theorems to correlated populations or samples; it involves an average entropy rate h instead of the entropy or relative entropy. We underline that maximum entropy prediction applies only to the empirical distribution or type, i.e. a probability distribution in the elementary state space $\mathcal{X}$. Its application to probability distributions describing another feature of the system, in a different space, leads to paradoxical results, as illustrated above.

Another example used by Boltzmann (Cover and Thomas 2006) is that of N dice thrown on the table such that the sum of the spots on their visible faces is some integer value $N\alpha$. We wish to known the most probable macrostate, where a macrostate describes the number of dice showing k spots for each $k = 1, \ldots, 6$. The answer is given by maximizing entropy of the probability distribution, namely $(p_k)_{k=1,\ldots,6}$ under the constraint $\sum_{k=1}^{6} kp_k = \alpha$ on its average (and the normalization constraint $\sum_{k=1}^{6} p_k = 1$). This yields $p_k = \exp(\lambda_0 + k\lambda_1)$ where λ_0 and λ_1 are chosen so as to verify the constraint of normalization and the constraint on the average. The rationale of using here maximum entropy principle comes from the theory of types and concentration theorems, § 3.1: For any fixed ϵ, we have the following asymptotic behavior for the conditional probability:

$$\lim_{N\to\infty} \text{Prob}\left\{ \bar{x}_N, |p_{\bar{x}_N} - p^*| < \epsilon \;\middle|\; \sum_{k=1}^{6} kp_{\bar{x}_N}(k) = \alpha_N \right\} = 1 \tag{37}$$

where α_N is a sequence tending to α when N tends to infinity (and such that $N\alpha_N$ is an integer), and p^*_α is the probability distribution on the elementary state space maximizing entropy at fixed average α. Prob corresponds to equiprobable configurations, namely it is given in the case of dice by the cardinal divided by the total number 6^N of configurations (uniform reference distribution).

In conclusion, maximum entropy applies safely only to the inference of the elementary distribution in a population of independent and identical individuals with discrete states. It supplements standard estimation method of a distribution from the empirical one (normalized histogram) by providing bounds on the convergence rate to the true distribution and controlling finite-sampling errors. Maximum entropy arguments can also be used to justify a *parametric expression* of the form (33) or (34) for a distribution. An ecological example, namely the reconstruction of species spatial distribution as a function of bioclimatic fields, can be found in (Phillips *et al.* 2006; Phillips and Dudík 2008). It extends into a maximum relative entropy principle when the issue is to update a prior distribution (reference distribution p_0) with additional knowledge and constraints, basically replacing entropy with relative entropy in the statements.

A corollary of Shannon maximum entropy principle is *Burg maximum entropy theorem*. It states

that the process wich maximizes entropy subject to correlation constraints is an appropriate autoregressive Gaussian process (Cover and Thomas 2006). More explicitly, the stochastic process maximizing the entropy rate given the correlations $\langle X_j X_{j+k}\rangle = \alpha_k$ for $k = 0, \ldots, K$ is the Kth-order Gauss-Markov process $X_j = -\sum_{k=1}^{K} a_k X_{j-k} + Z_j$ where Z_j is an independent and uncorrelated centered Gaussian process of variance σ^2 and the values of $(a_k)_{k=1,\ldots,K}$ and σ^2 are chosen so that the constraints are satisfied. A corollary is the fact that the entropy of a finite segment of a stochastic process is bounded above by the entropy of a segment of a Gaussian process with the same covariance structure (Cover and Thomas 2006). This theorem validates the use of autoregressive models as the less biased fit of data knowing only their correlations. Nevertheless, it does not validate an autoregressive model as an explanation of the underlying process. In the same spirit, the fact that the least biased fit of a distribution with a given mean μ and variance σ^2 is the normal distribution $\mathcal{N}(\mu, \sigma^2)$ does not prove that the distribution is indeed a Gaussian distribution. Here dedicated hypothesis testing should be developed in order to check whether the underlying process is indeed linear, or not.

3.5. *An illustration: types for uncorrelated random graphs*

Let us mention a possible extension of the method of types to uncorrelated random graphs, belonging to the recently expanding statistical mechanics of networks. A graph is fully specified by a set of N nodes and an adjacency matrix A describing the edges between these nodes (i.e. $A_{ij} = 1$ if there is an edge between the nodes i and j, else $A_{ij} = 0$, in particular $A_{ii} = 0$). Defining the degree k_i of the node i as the number of edges linked to i (explicitly, $k_i = \sum_{j=1}^{N} A_{ij}$), a degree sequence $[k](A)$ can be associated with the adjacency matrix A. We deduce the normalized histogram $p_A(k)$ of this sequence, which is nothing but the empirical degree distribution. The average degree $\langle k\rangle_{p_A}$ with respect to this empirical distribution p_A coicindes with the degree average $\sum_{i=1}^{N} k_i/N$, itself equal to $2M(A)/N$ where $M(A)$ is the number of edges of the graph. They are random variables insofar as A is itself a random variable when considering statistical ensembles of random graphs. A graph can be considered at four different hierarchical levels:
– the adjacency matrix $A \in \{0,1\}^{N^2}$, containing the full knowledge about the graph, at the level of the pair of nodes;
– the degree sequence $[k](A) \in \mathbf{N}^N$, at the node level, in which permutations (of the nodes) matter;
– the empirical degreee distribution $p_A(k) \in \mathcal{P}(\mathbf{N})$, which is invariant upon node permutations;
– the empirical average $2M(A)/N$ of the degrees, which coincides with the statistical average according to the distribution p_A.
At this stage, we can work with two different statistical ensembles of graphs:
— the microcanonical ensemble $\mathcal{E}^N_{micro}(M_0) = \{A, M(A) = M_0\}$ endowed with uniform probability distribution $\mathcal{Q}^N_{micro}(M_0) = 1/|\mathcal{E}^N_{micro}(M_0)|$ (some tolerance δM can relax the condition $M(A) = M_0$, with no quantitative consequence at the level of entropy in the limit $N \to \infty$)
— the canonical ensemble $\mathcal{E}^N$ endowed with the Gibbs probability distribution $\mathcal{Q}^N_{can}(M_0)$ satisfying the maximum entropy criterion under the constraint $\langle M\rangle_{\mathcal{Q}^N_{can}(M_0)} = M_0$.

Let us consider the case of an uncorrelated graph with degree distribution p_0, namely the N degrees are drawn at random, independently, according to the distribution p_0. The degree sequence $[k]$ is thus a realization of an uncorrelated and uniform sequence with distribution p^0, and it is distributed according to the product distribution $p_0^{\otimes N}$.The empirical degree distribution p_A can be seen as the type $p_{[k](A)}$ of the degree sequence $[k](A)$, in a way similar to the type of a random sequence in probability theory. We denote $\mathcal{N}_N(p)$ the number of sequences of length N having the type $p \in \mathcal{P}(\mathbf{N})$. Csiszár-Körner theorem then states (Csiszár and Körner 1981): for any sequence $(p_N)_N$ such that $\lim_{N\to\infty} p_N = p_0$,

$$\lim_{N\to\infty} (1/N) \log \mathcal{N}_N(p_N) = H(p_0) \tag{38}$$

and for any convex set $\mathcal{C} \subset \mathcal{P}(\mathbf{N})$, Sanov's large deviation theorem states that:

$$\lim_{N\to\infty} (1/N) \log p_0^{\otimes N}\{[k], p_{[k]} \in \mathcal{C})\} = -\inf_{p\in\mathcal{C}} D(p||p_0) \tag{39}$$

4. Shannon entropy rate

4.1. *Definition*

For a stationary stochastic process $(X_t)_{t\geq 0}$ (in discrete time t) with values in a finite set $\mathcal{X}$, Shannon entropy of the array $(X_1, \ldots, X_n)$ is termed *block entropy* of order n and denoted H_n. It is the Shannon entropy of the n-word distribution p_n, namely:

$$H_n \equiv -\sum_{\bar{w}_n} p_n(\bar{w}_n) \log_2 p_n(\bar{w}_n) = h(p_n) \tag{40}$$

where the sum runs over all the possible n-words $\bar{w}_n$. The n-block entropy captures quantitatively correlations of range shorter than n, by contrast with the simple entropy $H = H_1$ which is only sensitive to the frequencies of the different elementary states (also termed "symbols" henceforth). Considering n-words and associated block-entropy should not be confused with coarse-graining nor with local average, § 2.5. The latter take place in the state space of a single variable $\mathcal{X}$ while p_n is a probability distribution in $\mathcal{X}^n$. For a stationary process, it follows from the definition and properties of the conditional entropy that (Karlin and Taylor 1975):

$$0 \leq H(X_{n+1} \,|\, X_1, \ldots, X_n) \leq H(X_{n+1} \,|\, X_2, \ldots, X_n) = H(X_n \,|\, X_1, \ldots, X_{n-1}) \tag{41}$$

This inequality could be rewritten $0 \leq H_{n+1} - H_n \leq H_n - H_{n-1}$ from which follows the existence of the *Shannon entropy rate* (Karlin and Taylor 1975; Cover and Thomas 2006):

$$h = \lim_{n\to\infty} H_{n+1} - H_n = \lim_{n\to\infty} H(X_{n+1} \,|\, X_1, \ldots, X_n) = H(X_0 \mid \overleftarrow{X}) \tag{42}$$

where $\overleftarrow{X} = (X_i)_{-\infty<i\leq -1}$. This entropy rate h is equivalently defined as the limit (Karlin and Taylor 1975; Cover and Thomas 2006):

$$h = \lim_{n\to\infty} \frac{H_n}{n} \tag{43}$$

This latter limit exists as soon as $\lim_{n\to\infty} H_{n+1} - H_n$ exists, and it then takes the same value; we shall here consider situations where the two limits exist, hence coincide. h is an asymptotic quantity, characterizing the global statistical features of the source. In particular it captures correlations of any range. It thus provides a quantitative measure of the overall temporal organization of the process. Let us denote

$$h_n = H_{n+1} - H_n = H(X_{n+1}|X_1, \ldots, X_n) \quad \text{and} \quad h_{n,av} = \frac{H_n}{n} \tag{44}$$

These intermediate quantities are monotonously decreasing toward their common limit h, and thus provide upper bounds on the entropy rate according to

$$h_{n,av} \geq h_n \geq h = \lim_{n\to\infty} h_n = \lim_{n\to\infty} h_{n,av} \tag{45}$$

An important point in using entropy rate for data analysis (Lesne *et al.* 2009) is that h makes sense for both deterministic and stochastic sources. Considering a sequence $(X_1, \ldots, X_n)$ of length n, it can be shown (Karlin and Taylor 1975) that random shuffle σ increases entropy, namely $H_n(\sigma.X) \geq H_n(X)$ except for an uncorrelated stationary process, for which $H_n(\sigma.X) = H_n(X) = nH_1(X)$. This property is exploited in surrogate methods to assess that an experimental sequence is not produced by an uncorrelated stationary source. The argument relies on showing that its estimated entropy rate is significantly lower than most of the entropy rates estimated from the shuffled sequences. Entropy rate estimation from data and interpretation in practical contexts is a

whole domain of research, deserving on its own a critical review (Kantz and Schrieber 1997; Lesne *et al.* 2009), far beyond the scope of the present paper.

4.2. *Examples and special cases*

For a sequence of independent and identically distributed random variables, $h = H_1$, i.e. h reaches its upper bound (at given symbol frequencies). Temporal correlations always lower h. For a stationary Markov chain of order 1, $h = H_2 - H_1$, while for a stationary Markov chain of order q, $H_n = H_q + (n-q)h$ as soon as $n \geq q$. In this case, $h_n = h$ exactly as soon as $n \geq q$, whereas $h_{n,av}$ gives only an approximation of h, with a remaining positive term $[H_q - qh]/n$. Accordingly, in the general case, h_n is the entropy rate of the Markov approximation of order n of the source. Note that the entropy rate of a first-order Markov chain with transition matrix

$$M(p) = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \tag{46}$$

equals that of a Bernoulli process B with probability $\mathrm{Prob}(B = 1) = p$ (both entropy rates are equal to $h = -p\log_2 p - (1-p)\log_2(1-p)$). This illustrates the fact that there is no one-to-one correspondence between entropy rates and processes. There is no way to directly infer insights on the underlying process and its features from the value of h itself: only a differential study makes sense. It could rely on the comparison between either two experimental systems (for classification purposes), or a real system and a model (for quality assessment), or two models (for model selection purposes). Only $h = H_1$ is directly meaningful, indicating the absence of temporal correlations in the process. Even $h = 0$ does not lead to a clear-cut conclusion since it is observed for both a periodic dynamics and a dynamics at the onset of chaos.

Let us also consider a less simple situation, that of a *hidden Markov model* $Y_t = X_t \oplus E_t$ where $\oplus$ is the exclusive-or logical operation, X is a binary Markov chain with a symmetric transition matrix (46) as above, and E is a Bernoulli noise process with parameter ϵ. Their common entropy rate is $h(E) = H_1(E) = -\epsilon\log_2\epsilon - (1-\epsilon)\log_2(1-\epsilon)$. In other words, Y_t is obtained by randomly flipping the symbol X_t with a probability ϵ, without any correlation between the successive flips. This is the typical case of a noisy transmission, where Y is the outcome of a noisy channel fed with an input X. It can be shown (Zuk *et al.* 2005) for small ϵ that $h(Y) = h(X) + \epsilon c_1 + \mathcal{O}(\epsilon^2)$ with $c_1 = 2(1-2p)\log_2[(1-p)/p]$. It is to note that $c_1 > 0$ as soon as $0 < p < 1$ hence $h(Y) > h(X)$ at least for small enough noise. Observing Y_t is associated with a greater surprise than observing X_t since an additional degree of randomness, the noise E_t, sets in. Using the fact that $H_n(Y) \geq h(Y)$, it follows that for δ small enough and n_0 large enough, the inequality $H_n(Y) > \delta + H_n(X)$ holds for any $n \geq n_0$. It is also to be underlined that $H_1(X) = H_1(Y) = 1$ (the noise does not break the symmetry insofar as the stationary state of both processes X and Y corresponds to equiprobability between symbols 0 and 1). Accordingly the difference between the input and output in terms of information contents and the influence of noise on transmission cannot be appreciated using only Shannon entropy.

4.3. *Information-theoretic interpretation*

In the information-theoretic interpretation of Shannon entropy H, § 2.2, the inputs were the elementary states or symbols $x \in \mathcal{X}$. Let us now consider the more realistic case where the message is formed by the *concatenation of symbols* emitted at each time by the source. For independent symbols, the source is still fully characterized by the entropy H of the elementary distribution. In the general case, time correlations are present between the sucessive symbols, and one has recourse to h to characterize the source. It is thus important to distinguish H and h: $h \leq H \leq \log_2|\mathcal{X}|$ and $h < H$ as soon as correlations are present. Indeed, using the stationarity of the process (Feldman

2002):

$$\begin{aligned} h &= \lim_{N\to\infty} H(X_0|X_{-1},\ldots,X_{1-N}) \\ &= H(X_0) - \lim_{N\to\infty} I(X_0\ ;\ X_{-1},\ldots,X_{1-N}) \\ &= H(X_0) - I(X_0;\overleftarrow{X}) \end{aligned} \tag{47}$$

from which we deduce that for a stationary source, $h = H_1$ if and only if there are no correlations between X_0 and $\overleftarrow{X}$. The entropy rate h captures both the unevenness of the symbol distribution and the correlations along the sequence, in a non additive way so that it is impossible to disentangle the two contributions. By contrast, H provides only quantitative characterization of the unevenness of the probability distribution of the symbols. Using the expression $h = \lim_{n\to\infty} H(X_0|X_{-1},\ldots,X_{-n})$, another interpretation of h is the information required to predicted X_0 knowing the whole past.

From its very definition and the information-theoretic interpretation of Shannon entropy, § 2.2, h is the average information brought by the observation of an additional symbol (Feldman 2002). Equivalently, the average missing information to predict the value of the next symbol in $\mathcal{X}$ is not $\log_2|\mathcal{X}|$ bits (1 bit for a binary sequence) but h bits. Indeed, some knowledge is brought by both the time correlations and the unevenness of the symbols frequency distribution. It means that some redundancy is present in a sequence of length N and, on the average, $N_{eff} = Nh/\log_2|\mathcal{X}|$ bits are enough to represent the sequence ($N_{eff} = Nh$ in case of a binary sequence).

Entropy rate also plays a role in statistics, insofar as it captures time correlations of the process, which centrally control error bars in estimation issues. For instance, for a stationary Gaussian process X, it can be shown (Cover and Thomas 2006) that the variance σ^2_∞ of the error of the best estimate of X_n given the infinite past is related to the entropy rate $h(X)$ of the process according to $2\pi e\sigma^2_\infty = 2^{2h(X)}$. More generally, in the issue of estimation from an experimental sequence, $N_{eff} = Nh(X)/\log_2|\mathcal{X}|$ is the effective length of the sequence, relevant for appreciating the importance of finite-size effects. The notions of entropy rate $H(X)$ and effective length N_{eff} thus provide the foundation of estimation theorems for correlated samples, e.g. the estimation of the underlying distribution from the observation of a time-correlated trajectory (Sokal and Thomas 1989; Sokal 1997; Lesne *et al.* 2009).

4.4. *Derived notions*

There is still a wide range of possible temporal structures for a dynamics characterized by a given entropy rate h. This observation motivated the search for additional measures to quantitatively characterize temporal structures or patterns and their statistical complexity (Feldman and Crutchfield 1998; Feldman 2002). A first direction has been to consider a quadratic function $Q(p) = (H(p)/H_{max})[1-H(p)/H_{max}]$ where $H_{max} = \log_2|\mathcal{X}|$ is the maximum entropy observed for distributions on the same space $\mathcal{X}$ as p. The idea is to enforce the expected behavior of a statistical measure of complexity, namely vanishing for regular, e.g. periodic, and fully random distributions (Shinner *et al.* 1999). Nevertheless, this quantity $Q(p)$ contains almost exactly the same knowledge about the distribution p than the entropy $H(p)$. It describes its features almost exactly at the same level and in the same way, as shown by the inverse formula $H(p)/H_{max} = [1 \pm \sqrt{1-4Q}]/2$. In fact, Q contains slightly less information since H and $H - H_{max}$ correspond to the same value of Q, meaning that an additional degeneracy is introduced in $p \to Q(p)$ compared to $p \to H(p)$. Similarly, the quantity $h(h_{max} - h)$ is not a complexity measure since it does not bring further insights on the structure and organization of the system compared to entropy rate h (it is not enough that it vanishes for regular and fully random sources). A more insightful notion is the *effective measure complexity* (Grassberger 1986; Gell-Mann and Lloyd 1996; Gell-Mann and Lloyd

2003), also termed *excess entropy* (Feldman 2002):

$$E = I(\overleftarrow{X} \mid \overrightarrow{X}) = I(x_{-\infty}^{-1} ; x_0^{+\infty}) \tag{48}$$

For instance, $h = 0$ is observed in several very different cases, e.g. periodic signals and onset of chaos. Excess entropy allows to discriminate the different situations associated with a vanishing entrppy by capturing the way H_n/n converges to $h = 0$. For instance, $H_n = const$ for a periodic signal while $H_n \sim \log_2 n$ at the onset of chaos (Feldman 2002). More generally (Grassberger 1986) $H_n \sim E + nh + \text{h.o.}$. Excess entropy equivalently expresses (Badii and Politi 1997):

$$E = \lim_{n\to\infty} (H_n - nh) = \lim_{n\to\infty} \frac{2H_n - H_{2n}}{n} = \sum_{n=1}^{\infty} n(h_{n-1} - h_n) + H_1 - h \tag{49}$$

A natural extension of the entropy rate is the *mutual information rate* (Gray 1990; Blanc *et al.* 2008)

$$i(X;Y) = \lim_{n\to\infty} (1/n) I([X_1, \ldots, X_n]; [Y_1, \ldots, Y_n]) \tag{50}$$

Denoting $\theta.X$ the shifted sequence, such that $(\theta.X)_t = X_{t+\theta}$, it is possible to show (Blanc *et al.* 2011) that the mutual information rate satisfies $i(X, \theta.X) = h(X)$ and $i(X, \theta.Y) = i(X, Y)$. Shannon in its historical paper of 1948, in the section devoted to transmission in a noisy channel, actually introduced the mutual information rate between the input signal and the output signal, and named it the *rate of actual transmission*. Denoting X the input (message emitted by the source) and Y the output (message after transmission in the channel or more generally any input-output device), the conditional entropy rate $h(X|Y) = h(X, Y) - h(Y)$ measures the average ambiguity of the output signal, i.e. the entropy of the message X emitted by the source given the output Y. One has $h(X|Y) = 0$ as soon as the knowledge of the output sequence $(y_1, \ldots, y_N)$ allows to determine the input message. In other words, $h(X|Y)$ is the amount of additional information that must be supplied per unit time to correct the transmitted message Y and recover X, while $h(Y|X)$ is the part due to noise in $h(Y)$. These two quantities are directly related to the mutual information rate as follows:

$$i(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \tag{51}$$

A different extension is based on the introduction of the *Rényi entropy* of order α (Rached *et al.* 2001):

$$H^{(\alpha)}(p) \equiv \frac{1}{1-\alpha} \log_2 \left(\sum_i p_i^\alpha \right) \tag{52}$$

and Rényi divergence of order α

$$D^{(\alpha)}(p||q) \equiv \frac{1}{\alpha - 1} \log_2 \left(\sum_i p_i^\alpha q_i^{1-\alpha} \right) \tag{53}$$

recovering respectively Shannon entropy and Kullback-Leibler divergence for $\alpha = 1$, according to $\lim_{\alpha\to 1} H^{(\alpha)}(p) = H(p)$ and $\lim_{\alpha\to 1} D^{(\alpha)}(p||q) = D(p||q)$. Similarly, one could extend the Rényi entropy of order α into an entropy rate $h^{(\alpha)} = \lim_{n\to\infty} H^{(\alpha)}(p_n)/n$ where p_n is the n-word distribution of the source. The rationale for considering those extended entropy rates is to give a tunable weight to rare or frequent events, eg. rare events contribute relative more in $h^{(\alpha)}$ for $\alpha = 1$ than in h. Their drawback is the lack of subadditivity property except for $\alpha = 0$ and $\alpha = 1$. It also flaws non extensive thermostatistics which is based on these generalized entropies (Balian 2004; Balian 2005).

4.5. *Spatial extension*

Block-entropies and entropy rate are statistical descriptors of a time series. Only technical care is needed to extend these notions and apply them to the quantitification of a *spatially extended structure* by considering spatial labels (x, y) or (x, y, z) instead of time t. One can obviously compute the Shannon entropy of the probability distribution describing the fraction of space occupied by the different species forming the pattern (normalized over the different species). A more refined quantification involves Shannon entropies at different scales of observation ϵ (local averages). Considering a partition of the space in $N(\epsilon)$ disjoint cells of size ϵ and denoting $p_i(\epsilon)$ the measure of the cell i (with $\sum_{i=1}^{N(\epsilon)} p_i(\epsilon) = 1$, normalized over the different spatial cells), a meaningful index is the *information dimension* describing the ϵ-dependence of the entropy (Badii and Politi 1997; Castiglione *et al.* 2008):

$$D = \lim_{\epsilon \to 0} \frac{\sum_{i=1}^{N(\epsilon)} p_i(\epsilon) \log_2 p_i(\epsilon)}{\log_2 \epsilon} \tag{54}$$

Another promising but yet not much developed index is the multivariate extension of the entropy rate devised for a time sequence. One has to consider increasing sequences $(B_n)_n$ of multidimensional blocks (e.g. squares for a spatial structure in the plane) and $h = \lim_{n\to\infty} H_{B_n}/|B_n|$, then check that the resulting entropy rate does not depend on the chosen sequence (i.e. if $An \subset B_n \subset A_{n+1}$, then $h([A]) = h([B])$). It would allow to evidence the existence of patterns, exactly like the entropy rate h is exploited in time series analysis to evidence non trivial temporal organization (Kantz and Schrieber 1997; Lesne *et al.* 2009). In this context, one should not confuse:

(i) the quantification of spatial structures by means of statistical entropy;

(ii) the investigation of thermodynamic entropy production in dissipative spatial structures.

To my knowledge, whether a relationship exists or not between the degree of spatial order of the structure and pattern and the thermodynamic entropy production when this structure or pattern is associated to nonequilibrium state of some process (dissipative structure) is still an open question (Mahara and Yamaguchi 2010). Presumably there is no universal link since entropy production directly involves the dynamics of the system while the pattern statistical entropy quantifies only the stationary outcome of the dynamics. In particular, different dynamics (with different entropy productions) could produce the same stationary patterns hence be associated with the same statistical entropy.

5. Asymptotic theorems and global behavior of correlated sequences

5.1. *Shannon-McMillan-Breiman theorem*

An extension of concentrations theorems to the case of correlated sequences is provided by *Shannon-McMillan-Breiman theorem* (already stated for Markov chains in (Shannon 1948), Theorem 3). Under an assumption of stationarity and ergodicity of the considered stochastic process, this theorem states that the number of typical m-words (i.e. that have the same properties corresponding to almost sure behavior) behaves like e^{mh} as $m \to \infty$ where the exponent h is the entropy rate of the source (Shannon 1948; McMillan 1953; Breiman 1957; Cover and Thomas 2006) . A corollary of this theorem is the *Asymptotic equipartition property*, stating that the probability $p_m(\bar{w}_m)$ of a typical m-word $\bar{w}_m$ takes asymptotically the value e^{-mh} common to all typical m-words, hence the name "equipartition". The statement has to be made more rigorous since the limiting behavior of the probabilities when $m \to \infty$ is still a function of m. Introducing the random variables $\hat{P}_m$ (depending on the whole realization $\bar{x}$ of the symbolic sequence) such that $\hat{P}_m(\bar{x}) = p_m(x_0, \ldots, x_{m-1})$, the asymptotic equipartition property writes:

$$\lim_{m\to\infty} (-1/m) \log_2 \hat{P}_m \to h \qquad \text{in probability} \tag{55}$$

i.e. for any $\delta > 0$ and $\epsilon > 0$ (arbitrary small), there exists a word-size threshold $m^*(\delta, \epsilon)$ such that:

$\text{Prob}(\{\bar{x}, p_m(x_0, \ldots, x_{m-1}) > 2^{m(-h+\delta)}\}) < \epsilon$ and $\text{Prob}(\{\bar{x}, p_m(x_0, \ldots, x_{m-1}) < 2^{m(-h-\delta)}\}) < \epsilon$

for any $m \geq m^*(\delta, \epsilon)$, or equivalently, in terms of m-word subset:

$p_m(\{\bar{w}_m, p_m(\bar{w}_m) > 2^{m(-h+\delta)}\}) < \epsilon$ and $p_m(\{\bar{w}_m, p_m(\bar{w}_m) < 2^{m(-h-\delta)}\}) < \epsilon$.

Asymptotic equipartition property for a sequence of independent and identically distributed variables is a mere consequence of the law of large numbers, stating that $(-1/N)\sum_{i=1}^{N} \log_2[p(X_i)]$ converges to $\langle \log_2[p(X)] \rangle = H(p)$ for N tending to infinity. Shannon-McMillan-Breiman theorem extends the law to correlated sequences. Nevertheless, all available results apply to stationary sources, which could be a strong limitation in practical situations.

Another corollary of Shannon-McMillan-Breiman theorem is to describe quantitatively how h accounts in an effective way for the correlations present within the sequence. Namely, the effective probability of a new symbol, knowing the sequence of length l that precedes, is asymptotically (i.e. for $l \to \infty$) either e^{-h} or 0 whether the ensuing $(l+1)$-word is typical or not. By contrast, it is equal to the symbol frequency in the case when there is no correlations within the sequence. We recover the interpretation of h as being the average information brought by the observation of an additional symbol. A pedagogical proof is given in (Algoet and Cover 1988) and (Karlin and Taylor 1975).

For N asymptotically large, Shannon-McMillan-Breiman theorem ensures that, up to second-order terms, one has $H_N \approx \log_2 \mathcal{N}_N$ where $\mathcal{N}_N$ is the number of (asymptotically equiprobable) typical sequences. We shall see, § 8.2, that this approximate formulation of the Shannon-McMillan-Breiman theorem parallels the definition of Boltzmann entropy in the microcanonical ensemble. Here we can interpret H_N as the average information brought by the reception of a message (i.e. one of these $\mathcal{N}_N$ messages). We emphasize that Shannon-McMillan-Breiman theorem deals with *probable* sequences, by contrast with a *grammar*, which describes the set of *possible* sequences, or equivalently the rules for generating all the possible sequences.

Two derived formulations of Shannon-McMillan-Breiman theorem can be useful. Let $\mathcal{N}_N(\epsilon)$ be the cardinal of the smallest ensemble E of N-sequences whose total measure overwhelms $1 - \epsilon$. Then $\lim_{N\to\infty}(1/N)\log_2 \mathcal{N}_N(\epsilon) = h$. The second one is given as Theorem 4 in (Shannon 1948) and states as follows. Let us order the sequences of length N at decreasing probabilities and define $n(q)$ as the number of sequences (starting with the most probable one) needed to accumulate a total probability q ($0 < q < 1$ being fixed, independent of N). Then $\lim_{N\to\infty}(1/N)\log_2 n(q) = h$.

5.2. *Compression of a random source*

In this subsection, we address the issue of the compression of a random source. We deal here with *ensemble compression*, that is, how to transmit the most economically any one message from a given set. The question is to determine the minimal piece of knowledge that should be transmitted to faithfully discriminate one message from all the other possible ones. The reference to a specific ensemble of messages, or more generally of events, is essential. It is currently specified through a probability distribution. A special case is a source generating successive symbols, for which we deal with the compression of an *ensemble of sequences*. An essentially different issue is the compression of a *single* sequence, that will be addressed in § 6.2.

In the most general case, the optimal encoding for source compression has been introduced by Shannon and is known today as the *Shannon-Fano code*. Given a finite set $\mathcal{X}$ of elements x and their distribution of probability $p(x)$, a binary code is a correspondence $x \to w(x)$ where $w(x)$ is a binary word, i.e. a finite string of 0 and 1 representing x. We denote W the finite set of codewords representing the elements of $\mathcal{X}$ and l_w the length of the codeword w. The code is univoquely and

locally decodable if the following condition, known as the *Kraft's inequality*, is satisfied:

$$\Sigma(l) = \sum_{w \in W} 2^{-l_w} \leq 1 \tag{56}$$

The code is said to be *compact* if the inequality is replaced by an equality. Compact codes correspond to optimal codes insofar as their codewords have a minimal length. Otherwise, it is possible to compress the coding of the elements of $\mathcal{X}$ while preserving the univoquely and locally decodable character of the code: indeed, if $l'_w \geq l_w$ for any word w, then $\Sigma(l') \leq \Sigma(l)$. Given the probability $\tilde{p}(w) \equiv p[x(w)]$, minimization of the average length $\sum_{w \in W} l_w \tilde{p}_w$ at fixed $\Sigma(l) = 1$ yields

$$l_w = \log_2(1/\tilde{p}_w) \tag{57}$$

equal to the missing information required to specify $x(w)$. The average codeword length is then:

$$\bar{l} = -\sum_{w \in W} \tilde{p}(w - \log_2 \tilde{p}(w) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = I(p) \tag{58}$$

Shannon entropy thus gives the *average codeword length* for an optimal binary code, achieving the optimal compression of the information needed to represent events x knowing only their probability of occurrence $p(x)$ in the event set $\mathcal{X}$.

Kullback-Leibler divergence $D(p||q)$, § 2.3, measures the extra average length of the codewords when one uses an ill-adapted Shannon-Fano binary code, namely a code adapted to the probability distribution q whereas the actual probability is p (think for instance of a Morse code optimized for English used to transmit a french text). Indeed, the prescription of a Shannon-Fano code adapted to the probability distribution q is to use a codeword of length $l_x = -\log_2 q_x$ to represent an event $x \in \mathcal{X}$ of probability q_x. If the actual probability is p, the average length is $\langle l \rangle = -\sum_{x \in \mathcal{X}} p_x \log_2 q_x$ while the optimal average length is $\langle l \rangle_{opt} = -\sum_{x \in \mathcal{X}} p_x \log_2 p_x$, yielding $\langle l \rangle - \langle l \rangle_{opt} = D(p||q)$.

In the case of a sequence-generating source, the source X emits with the probability $p(\bar{x}_N)$ a message $\bar{x}_N$ of length N written in the alphabet $\mathcal{X}$ (i.e. the set of elementary events, usually termed "symbols"). For transmission or storage purposes, one encodes the message into a binary sequence $\bar{w}$, aiming moreover at minimizing the length of the coded message. A complete absence of knowledge at time t on the next symbol x_{t+1} generated by the source X is achieved in case of a sequence of independent and equiprobable symbols (uncorrelated and fully random source), in which case $h(X) = H_1 = \log_2 |\mathcal{X}|$. Accordingly, Shannon defined the redundancy of a source X as $1 - h(X)/\log_2[\mathcal{X}|$. Compression takes benefit of the discrepancies from such a full randomness, originating either in an uneven distribution of the elementary events (meaning that some symbols are observed more often), or in time correlations (meaning that observing a symbol x_t at time t somehow conditions the subsequent observation x_{t+1}), or both.

Source compression is already possible for a sequence of independent and identically distributed discrete random variables. It takes benefit of the unevenness of the frequencies of the different symbols: in the asymptotic limit $N \to \infty$, there are only $2^{NH(p)}$ typical sequences instead of the maximal number $2^{N|\mathcal{X}|}$ of possible sequences. More constructively, for each realization $(x_1, \ldots, x_N)$, there exists a binary string $\bar{w}_N(x_1, \ldots, x_N)$, given by a one-to-one mapping, and such that for any arbitrary small ϵ, there exists N_ϵ for which the average length satisfies $\langle |\bar{w}_N| \rangle / N \leq H(p) + \epsilon$ for any $N \geq N_\epsilon$. It means that on the average, a realization $\bar{x}_N$ of the original sequence is encoded by NH bits for N large enough.

For a correlated sequence-generating source X, compression also takes benefit of the temporal correlations and the compression rate is controlled by the entropy rate $h(X) < H(X)$. After an optimal compression of the source X, there shoud be no more redundancy in the compressed source W. In other words, there should be no bias in the symbol distribution and no correlations

between the successive symbols in the binary sequences $\bar{w}$, so that knowing $w_1, \ldots, w_t$ gives no clue on the next symbol w_{t+1}: the entropy rate of the source W is equal to 1, meaning that the average missing information per symbol in sequences $\bar{w}$ takes its maximal value, equal to 1. Such an optimal encoding is achieved by the above Shannon-Fano code (where events are now sequences of length N). According to this coding procedure, the length of the binary sequence $\bar{w}_{\bar{x}_N}$ encoding the sequence $\bar{x}_N$ is $l(\bar{w}_{\bar{x}_N}) \sim \log_2(1/p(\bar{x}_N)$. Asymptotically, $\lim_{N\to\infty}(1/N)\langle l(\bar{w}_{\bar{x}})\rangle = h(X)$, from which follows that $h(X)$ is the average number of bits required to optimally encode a symbol emitted by the source. A realization of the original sequence being on the average encoded by $Nh(X)$ bits, the shortest binary sequences representing faithfully the original source during a duration N will have an average length $Nh(X)$. We see here that compression of sequences (with no loss of information) is controlled in the asymptotic limit $N \to \infty$, where only typical sequences (in the sense of Shannon-McMillan-Breiman theorem) are to be taken into account. In this limit, the entropy rate $h(X)$ gives a lower bound (per symbol of the original sequence) on the compression that could be achieved.

6. Relation to algorithmic complexity

6.1. *Kolmogorov complexity*

Algorithmic complexity, also termed *Kolmogorov complexity*, as been introduced independently by Chaitin, Kolmogorov and Solomonoff (Chaitin 1966; Solomonoff 1978; Durand and Zvonkine 2007) to characterize a single sequence. The notion of (average) missing information introduced in a probabilistic setting is replaced in the algorithmic approach by the length of the shortest program, with no reference to an ensemble of sequences nor an a priori probability. More explicitely, a single sequence $\bar{x}_N$ of length N can be compressed into a binary sequence of minimal length $K(\bar{x}_N)$ describing the shortest program able to generate $\bar{x}_N$. Any other program generating the sequence has a length $L(\bar{x}_N)$ such that $K(\bar{x}_N) \leq L(\bar{x}_N) \leq N$, i.e. $K(\bar{x}_N)$ provides a lower bound. For arbitrarily long sequences, the computer-specific contribution to $K(\bar{x}_N)$ becomes relatively negligible (Durand and Zvonkine 2007), hence algorithmic complexity is generally defined in reference to a universal Turing machine. Since entropy is not a measure of complexity, the name of algorithmic complexity is misleading: it is of little help to characterize the complexity of a source (Gell-Mann and Lloyd 1996; Gell-Mann and Lloyd 2003). It should rather be termed "algorithmic information" as recommended in (Badii and Politi 1997). The theory is of limited power for short sequences, but now stationarity of the sequence is not mandatory. The main issue with algorithmic complexity and its conditional extensions is their *non computability*. It leads to have recourse to lossless compression algorithms to approximate (upper bound) algorithmic complexity, see § 6.2.

In the same way as one defines an entropy rate h, one could consider an algorithmic information density for a given sequence $\bar{x}$, defined as $C(\bar{x}) = \lim_{n\to\infty} K(\bar{x}_n)/n$ where $\bar{x}_n$ is a n-word extracted from $\bar{x}$ (the limit does not depend on the choice of this block) (Badii and Politi 1997). Remarkably its has been shown (Ziv and Lempel 1978) that for a stationary and ergodic source, $C(\bar{x})$ coincides up to a normalization factor with the entropy rate of the source for almost all sequences $\bar{x}$ (typical sequences in the sense of Shannon-McMillan-Breiman theorem), namely $C(\bar{x}) = h/\log_2|\mathcal{X}|$. It follows that on the average,

$$\lim_{N\to\infty} \frac{\langle K(\bar{x}_N)\rangle}{N} = h/\log_2|\mathcal{X}| \tag{59}$$

The average-case growth rate of Kolmogorov complexity is thus related to Shannon entropy rate according to $\langle K(\bar{x}_N)\rangle \sim h/\log_2|\mathcal{X}|$ for N large enough (Feldman and Crutchfield 1998). The notion fruitfully extends to that of *conditional Kolmogorov complexity* $K(x|y)$ of a sequence or random object x. It describes the length of the shortest program able to generate x when making use of

extra knowledge y, for instance $K(x|A)$ knowing $x \in A$ (Gell-Mann and Lloyd 1996; Gell-Mann and Lloyd 2003).

6.2. *Lempel-Ziv compression scheme and coding theorems*

We already mentioned in § 5.2 that two compression issues are to be carefully distinguished: compression of an ensemble of sequence (source compression) and compression of a single sequence. In both case, compression is achieved by the most efficient/economical encoding, hence coding and compression issues are thus solved jointly, and are limited by the same bounds, involving (Csiszár and Körner 1981; Badii and Politi 1997; Falcioni *et al.* 2003):

– either the Shannon entropy rate for compression of a source of known probability, i.e. compression of the set of sequences emitted by the source (§ 5.2);

– or the algorithmic complexity for a single sequence emitted by an unknown source.

We focus on the second situation. Compression of a single sequence $\bar{x}$ is possible if $K(\bar{x}) < |\bar{x}|$ (accordingly, a sequence $\bar{x}$ with $K(\bar{x}) = |\bar{x}|$ is termed *incompressible*).

The difficulty for practical purposes, in which we have to consider finite-length sequences $\bar{x}_N$, is the incomputability of $\mathcal{K}(\bar{x}_N)$. In order to circumvent this incomputability, compression algorithms with no loss of information can be used to obtain upper bounds, the better the more efficient the compression is. One of the most successful is *Lempel-Ziv algorithm* (a variant is used today in JPEG compression software). Its general principle is to enumerate new substrings discovered as the sequence evolves from left to right (Badii and Politi 1997; Cover and Thomas 2006). According to the Lempel-Ziv scheme, the sequence of length N is parsed in $\mathcal{N}_w$ words. Two different parsings have been proposed, either (Lempel and Ziv 1976):

1 • 0 • 01 • 10 • 11 • 100 • 101 • 00 • 010 • 11...

delineating as a new word the shortest one that has not yet been encountered, or (Ziv and Lempel 1977):

1 • 0 • 01 • 101 • 1100 • 1010 • 001011 • ...

delineating as a new word the shortest subsequence that has not yet been encountered (Ziv and Lempel 1977) (the fourth word in the above example is thus 101 and not the 2-sequence 10 since the latter has already been seen). The parsing allows to encode efficiently the original sequence with no loss of information. Indeed, each new word appearing in the parsing is uniquely specified by the already encountered word by which it begins and the additional symbol by which it is completed. One then computes:

$$\hat{L}_0 = \frac{\mathcal{N}_w[1 + \log_k \mathcal{N}_w]}{N} \tag{60}$$

providing an upper bound on the algorithmic complexity rate of the original sequence.

A remarkable result for a stationary and ergodic source is *Lempel-Ziv theorem* (Ziv and Lempel 1978). It states that both algorithmic complexity rate and Lempel-Ziv complexity rate are asymptotically equal to h (up to a constant normalization factor depending on the definitions and choice of the logarithm base) for almost all sequences:

$$\lim_{N\to\infty} \frac{K(\bar{x}_N)}{N} = \lim_{N\to\infty} \hat{L}_0(\bar{x}_N) = \frac{h}{\ln k} \tag{61}$$

It means that for N large enough, $\hat{L}_0(\bar{x}_N)$ gives not only an upper bound on $K(\bar{x}_N)/N$ but also an approximation $\hat{L}_0(\bar{x}_N) \approx K(\bar{x}_N)/N$ with asymptotic equality. Lempel-Ziv theorem also means that almost all symbolic sequences have the same compressibility features, hence the computation can be equivalently performed with any typical sequence. From Shannon-McMillan-Breiman

theorem, § 5.1, typical sequences have a full measure, hence sequences drawn at random or observed experimentally are typical; only sequences generated from a specially chosen non generic initial condition might happen to be non typical. Hence, in practice, computing the Lempel-Ziv complexity $\hat{L}_0$ gives an estimate of the entropy rate h, up to a normalization factor and provided a sufficient convergence (N enough large) is achieved (Lesne *et al.* 2009). A simpler computation involves

$$\hat{L} = \frac{\mathcal{N}_w \log_2 N}{N} \quad \text{with} \quad \lim_{N\to\infty} \hat{L} = h \tag{62}$$

Replacing $\log_k$ by $\log_2$ makes the limit directly comparable to h whereas the original definition is normalized with a common upper bound equal to 1. Several variants and improvements of the original Lempel-Ziv algorithms have been developed, see for instance (Wyner and Ziv 1989).

6.3. *Randomness of a sequence*

One of the first formalization of the randomness of a sequence is due to von Mises: a single binary sequence is random if the limiting frequency of 1 exists and does not change when considering an infinite subsequence chosen at random (i.e. chosing the subset of labels without involving the actual values 0 or 1 associated to each one). Kolmogorov refined this notion into that of (N, ϵ)-randomness, of relevance for finite sequences of length N, with a fixed tolerance ϵ (see (Vovk and Shafer 2003) for a detailed historical account and original references).

Nowadays, algorithmic complexity theory gives a rigorous basis to what is randomness of a sequence (Falcioni *et al.* 2003; Parisi 2003; Castiglione *et al.* 2008). An incompressible sequence $\bar{x}$, i.e. such that $K(\bar{x}) = |\bar{x}|$, is termed *algorithmically random* (Li and Vitanyi 1997; Durand and Zvonkine 2007). This notion of randomness is stronger than statistical randomness since some statistically random sequences (whose digits pass the statistical of being uniformly distributed, e.g. the decimal digits of π), are not algorithmically random. It has been applied to real numbers by Martin-Löf (Martin-Löf 1966) : introducing a dyadic representation of real numbers, he proved that almost all binary sequences thus obtained (for the Lebesgue measure on the associated set of real numbers) have a maximal algorithmic complexity $C = 1$.

7. Relation to ergodic theory of dynamical systems

7.1. *Metric entropy*

Shannon entropy for discrete-valued and discrete-time stochastic process has an exact analog in ergodic theory of dynamical systems. It has been developed by Kolmogorov (Kolmogorov 196) and Sinai (Sinai 1959), and there called metric entropy or Kolmogorov-Sinai entropy. Given a discrete-time evolution $x_{n+1} = f(x_n)$ on a compact topological space, one considers a finite partition $\mathcal{P}_0$ and the refinements generated in the course of time by the map f, namely $\mathcal{P}_n = \mathcal{P}_0 \vee f^{-1}(\mathcal{P}_0) \vee \ldots \vee f^{-n}(\mathcal{P}_0)$. One then computes:

$$\widetilde{h}_n(\mathcal{P}_0) = -\frac{1}{n} \sum_{A_n \in \mathcal{P}_n} m(A_n) \ln m(A_n) \tag{63}$$

where m is the invariant measure under the action of f, and:

$$\widetilde{h}(\mathcal{P}_0) = \lim_n \widetilde{h}_n(\mathcal{P}_0) = \widetilde{h}(\mathcal{X}_0) \tag{64}$$

The *metric entropy*, or *Kolmogorov-Sinai entropy* is finally obtained as (Wehrl 1978):

$$h_{KS} = \sup_{\mathcal{P}_0} \widetilde{h}(\mathcal{X}_0) \tag{65}$$

It is actually a *rate* of entropy. Note that one uses ln instead of $\log_2$ in dynamical systems theory. This is a purely conventional choice, motivated by practical and historical reasons since the two quantities are related by a factor of $\ln 2$, namely $h_{KS} = h\,.\,\ln 2$ (i.e. $e^{h_{KS}} = 2^h$). Metric entropy h_{KS} has been introduced to solve the "isomorphism problem", i.e. whether there is a mapping between two seemingly different dynamical systems, preserving the dynamical and statistical relationships betwen the successive states. h_{KS} being invariant under any isomorphism, two dynamical systems with different values for h_{KS} are non isomorphic. It also proved to be a very fruitful quantity for quantifying the seemingly erratic and irregular behavior of chaotic dynamical systems (Kantz and Schrieber 1997). In some cases, e.g. one-dimensional Anosov maps, there exists partitions $\mathcal{P}$, termed *generating partitions*, such that the continuous dynamics is exactly isomorphic to a discrete stochastic process. It is then enough to know at each time the location of the trajectory in $\mathcal{P}$ (i.e. the symbol labeling at each time the corresponding element of the partition) to uniquely specify the initial condition in the continuous phase space, and to reconstruct the continuous trajectory from the symbolic sequence. For a generating partition $\mathcal{P}$, $\widetilde{h}(\mathcal{P})$ reaches its maximum value h_{KS}, and coincides up to a factor $\ln 2$ to the Shannon entropy rate of the symbolic sequence, namely $h_{KS} = h/\ln 2$. Accordingly metric entropy can be equally computed on discretized trajectories. One could say more: according to *Jewett-Krieger theorem* (Krieger 1970; Krieger 1972; Falcioni *et al.* 2003; Glasner 2003), a continuous-valued dynamical system in discrete time, with finite entropy $h_{KS} > 0$, is equivalent to a stochastic process with a finite number of states, and the minimal number m of states satisfies $e^{h_{KS}} \leq m < 1 + e^{h_{KS}}$. These results are the basis and justification of *symbolic dynamics*, replacing the analysis of the dynamical system generated by a continuous map by that of the symbolic sequences describing the evolution at the level of the generating partition $\mathcal{P}$.

For a deterministic dynamics, a positive metric entropy $h_{KS} > 0$ is currently considered as a criterion and quantitative index of *chaos* (Laguës and Lesne 2008; Castiglione *et al.* 2008). A justification is the relationship between h_{KS} and the sum of positive Lyapounov exponents (Pesin 1997)

$$h_{KS} \leq \sum_{i,\ \gamma_i \geq 0} \gamma_i = \sum_i \gamma_i^+ \tag{66}$$

This inequality, known as *Pesin inequality*, turns into an equality for sufficiently chaotic systems, e.g. Anosov systems (Ledrappier and Strelcyn 1982; Castiglione *et al.* 2008). It means that production of information (i.e. gain of information about initial conditions by observing the trajectory one step more) is provided only by unstable directions, hence $h_{KS} \leq \sum \gamma_i^+$. Negative Lyapunov exponents $\gamma_i < 0$, associated with stable directions, play here no role. The relevance of metric entropy in data analysis to globally quantify the temporal organisation of the evolution has been recognized in numerous applications. It is now a standard tool of nonlinear time series analysis, for both continuous-valued and discrete (symbolic) sequences (Kantz and Schrieber 1997).

One could define the so-called ϵ-*entropy* by considering a partition $\mathcal{P}_\epsilon$ of the phase space with cells of diameter bounded above by ϵ, instead of taking the supremum over all possible partitions as in (65). The noticeable point is that ϵ-entropy can be defined and computed for any dynamic process, whether deterministic or stochastic. A behavior $\lim_{\epsilon \to 0} h(\epsilon) = h_{KS}$ with $0 < h_{KS} < \infty$ is characteristic of a deterministic chaotic system. For a truly stochastic process, $h(\epsilon)$ diverges as ϵ tends to 0, and the form of its increase as a function of $1/\epsilon$ discriminates different kinds of stochastic processes, with trajectories all the more irregular as the increase is steeper. For instance, it behaves as $(1/\epsilon)^2$ for a Brownian process (Nicolis and Gaspard 1994; Falcioni *et al.* 2003; Castiglione *et al.* 2008).

7.2. *Topological entropy*

Another entropy-like quantity, *topological entropy* h_{top}, is of relevance for describing the overall statistical behavior of a dynamical system. Denoting $\mathcal{N}(n,\epsilon)$ the maximal number of trajectories ϵ-separated over n time steps (i.e. for at least one time between 0 and n, the distance between the trajectories is larger than ϵ), it is defined as :

$$h_{top} = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \log \mathcal{N}(n,\epsilon) \tag{67}$$

By contrast with metric entropy h which is relative to an invariant ergodic measure, h_{top} depends only on the distance endowing the phase space. It describes how many trajectories are required to span the phase space with a prescribed resolution. As for h, it is defined as a rate (entropy per unit time). When a generating partition exists and allow to investigate the statistical features of the dynamics on a reduced symbolic version, topological entropy expresses $\lim_{n\to\infty}(1/n)\log \mathcal{N}_n$ where $\mathcal{N}_n$ is the number of admissible n-words. This formula shows that topological entropy is in fact already present in Shannon's seminal paper. It coincides with the notion of *capacity* of a deterministic communication channel (Shannon 1948): $C = \lim_{N\to\infty}(1/N)\log_2 M_N$ where M_N is the number of signals of length N that could be transmitted in the channel.

The introduction of Rényi entropy rates, § 4.4, allows to unify metric and topological entropies in unique framework: one indeed recovers $h(\alpha = 1) = h_{KS}/\ln 2$ and $h(\alpha = 0) = h_{top}/\ln 2$. The generalized framework of Rényi entropies is of relevance in the application of thermodynamic formalism to dynamical systems and their multifractal analysis (Badii and Politi 1997), briefly presented in the next subsection.

7.3. *Thermodynamic formalism*

To derive in a systematic way all the relevant statistical features of a discrete-time dynamical system, an unifying framework has been developed. It is formally reminiscent of Boltzmann-Goibbs formalism in statistical mechanics and for this reason termed *thermodynamic formalism* (Ruelle 1978). The basic idea is to introduce the analog of a partition function, where the role of n-particles configurations is played by stretches of trajectories of duration n. Within a symbolic description of the dynamics, these stretches correspond to n-words, and the partition function writes (Badii and Politi 1997; Lesne 1998):

$$Z(n,q) \equiv \sum_{w_n} [p_n(w_n)]^q = \langle [p_n(w_n)]^{q-1} \rangle \tag{68}$$

The exponent $q-1$ can be seen as an inverse temperature.The relevant analog of free energy is:

$$I(n,q) = -\frac{\ln Z(n,q)}{(q-1)} \tag{69}$$

and its density:

$$J(q) = \lim_{n\to\infty} J(n,q) \quad \text{where} \quad J(n,q) = \frac{I(n,q)}{n} \tag{70}$$

It is straightforward to show that $J(q=1)$ coincides with the metric entropy h_{KS}. $I(n,q)$ is nothin but the Rényi entropy (§ 4.4) for n-words. The graph $q \to J(n,q)$ actually encapsulates the fluctuations of the local entropy $\kappa(w_n) = -(1/n)\ln p_n(w_n)$. The average $\sum_{w_n} \kappa(w_n)p(w_n)$ tends to h_{KS} as $n \to \infty$. Furthermore, Shannon-McMillan-Breiman theorem ensures that local entropies $\kappa(w_n)$ tend to h_{KS} almost surely (with respect to the ergodic invariant measure of relevance) as $n \to \infty$. A large deviation formulation:

$$e^{-N(q-1)J_q} = \int e^{-N[q\kappa - g(\kappa)]}\, d\kappa \tag{71}$$

yields to the following Legendre reciprocal transformations:

$$(q-1)J_q = \inf_{\kappa}[q\kappa - g(\kappa)] \quad (72)$$

$$g(\kappa) = \inf_{q}[q\kappa - (q-1)J_q] \quad (73)$$

The large deviation function $g(\kappa)$ is called the *entropy spectrum* (Badii and Politi 1997; Lesne 1998). It provides a full characterization of the local singularities of the ergodic invariant measure.

7.4. *Typicality, compressibility and predictibility*

Shannon-McMillan theorem can be re-formulated in the context of dynamical systems as the following asymptotic statement (Badii and Politi 1997). The probability $\mu(\epsilon, n, x_0)$ (with respect to the invariant ergodic measure μ) of finding an orbit remainin during n steps at a distance smaller than ϵ from the orbit of x_0 behaves as $\mu(\epsilon, n, x_0) \sim e^{D_1 + nh_{KS}}$ for μ-almost every x_0 in the limit as $n \to \infty$.

On the other hand, Brin and Katok proved a kind of topological version of Shannon-McMillan-Breiman theorem for any dynamical system with ergodic invariant measure μ. It states that $\lim_{\epsilon \to 0} \limsup_{n\to\infty} (1/n) \log \mu[B(x, n, \epsilon)] = h(\mu)$ where $B(x, n, \epsilon)$ is the set of initial conditions whose orbit remains during n steps at a distance lower than ϵ from the orbit of x (Brin and Katok 1983). It means that the probability (with respect to the invariant measure μ) that two trajectories stay close together during n steps decays exponentially with n.

The relationship between Shannon entropy rate and metric entropy reinforces the relevance of a probabilistic description of chaotic (deterministic) dynamical systems, i.e. the use of statistical descriptors to quantify their apparent randomness (Nicolis and Gaspard 1994). In particular, it gives another interpretation to their unpredictability, relating it to the incompressibility of the source and the high algorithmic complexity of μ-almost all trajectories, when encoded using a generating partition (Castiglione *et al.* 2008).

Beyond the above global view considering a dynamical system as a random source, one could also consider each trajectory isolatedly and compute the algorithmic complexity of its symbolic description. In this context, a theorem by Brudno and White assesses that for an autonomous ergodic dynamical system, the Kolmogorov complexity of allmost all trajectories (almost all with respect to the invariant ergodic measure) is equal to h_{KS} up to a constant normalization factor (Brudno 1983; White 1993; Castiglione *et al.* 2008) . This theorem is nothing but the deterministic version of Lempel-Ziv theorem (Ziv and Lempel 1978). Accordingly, for a symbolic trajectory, we have equivalence between being unpredictable, incompressible and algorithmic complex (Falcioni *et al.* 2003).

8. Relation to statistical physics

8.1. *The second principle of thermodynamics*

Thermodynamic entropy has been introduced by Clausius in 1865 (Clausius 1865). He postulated that there exists a state function S_{th} defined for equilibrium states, such that $\Delta S_{th}(AB) \equiv S_{th}(B) - S_{th}(A) = \int_A^B \delta Q / T_{source}$ where δQ is the quantity of heat exchanged between the system and external sources at temperature T_{source} along an arbitrary transformation of the system from the equilibrium state A and the equilibrium state B. The variation $\Delta S_{th}(AB)$ does not depend on the transformation but only on the initial and final states since S_{th} is assumed to be a state function (Gallavotti 2006). Equality hold if and only if the transformation is reversible. For isolated systems, more precisely thermodynamically closed systems for which $\delta Q = 0$, one has $\Delta S_{th} \geq 0$.

This statement is known as the *Second Principle* of thermodynamics. It is an empirical principle discriminating the phenomena that could occur from those that are thermodynamically forbidden. It is not expected to hold at molecular scale (Castiglione *et al.* 2008) and actually it does not. Recent advances describe the discrepancy to the second principle arising in small systems under the form of fluctuations theorems (Gallavotti 1998; Cohen and Gallavotti 1999; Evans and Searles 2002).

The second principle should not been confused with the *H-theorem* (Cercignani 1988b). The former is indeed a universal but empirical principle (presumed to be) valid for any thermodynamically closed macroscopic system whereas the latter is an exact (i.e. rigorously proved) property of the Boltzmann kinetic equation, whence the name of theorem (Castiglione *et al.* 2008). This equation, central to the kinetic theory of dilute gases, describes the evolution of the one-particle probability distribution function $f(\vec{r},\vec{v},t)$ of the gas within the framework of continuous-medium approximation. The H-theorem then states that the quantity $H_B = \int f(\vec{r},\vec{v},t)\ln f(\vec{r},\vec{v},t)d^3\vec{r}d^3\vec{v}$ can only decreases in the course of time. One has to write this quantity B in a discrete form $\sum_i f(\vec{r}_i,\vec{v}_i,t)\log f(\vec{r}_i,\vec{v}_i,t)\Delta^3\vec{r}\Delta^3\vec{v}$ where $\Delta^3\vec{r}\Delta^3\vec{v}$ is the elementary volume in the one-particle phase space to relate it to a Shannon entropy (Castiglione *et al.* 2008).The fact that H_B can only increase originates in the decorrelation approximation involved in the derivation of the Boltzmann equation (Cercignani 1988a). It amounts to replace 2-particle distributions arising in the interaction kernel by a product of one-particle distributions. H-theorem thus only indirectly and approximately describes a feature of the real world, insofar as the system behavior is properly accounted for by the Boltzmann kinetic equation. It should not be confused with a property of irreversibility of the real system.

8.2. *Boltzmann entropy and microcanonical ensemble*

In classical statistical mechanics, what is termed entropy is basically *Boltzmann entropy*, namely, a quantity related to the number Γ_N of N-particles microstates which have the same prescribed macroscopic properties:

$$S_B = k_B \ln \Gamma_N \tag{74}$$

where k_B is the Boltzmann constant $k_B = 1.38\,.\,10^{-23}$ J/K. This formula has been proposed by Boltzmann in 1877 (Boltzmann 1877; Cercignani 1988b; Castiglione *et al.* 2008) and it is written (in the form $S = \log W$) as an epitaph on his grave. Boltzmann had a discrete viewpoint, defining microstates as elementary volumes in the microscopic phase space (a space of dimension $6N$ if the system comprises N particles).

The starting point of the statistical description is usually the microcanonical ensemble (see (Castiglione *et al.* 2008) for a discussion of its relevance and validity). It corresponds to considering equiprobable microscopic configurations at fixed volume V and fixed energy U, with a tolerance δU. Boltzmann entropy is then proportional to the logarithm of the associated phase space volume $\Gamma(N,V,U,\delta U)$. It is to note that δU does not play any role in the thermodynamic limit $N \to \infty$. Since $\Gamma(N,V,U,\delta U)$ behaves as $\Gamma(N,V,U,\delta U) \sim \Gamma_1^N \delta U$, Boltzmann entropy is *extensive* (proportional to N) and the contribution of δU in $\ln\Gamma$ is a higher-order term in S_B that could be neglected for large N (namely $\ln \delta U$ is negligible compared to $N \ln \Gamma_1$ for large N). Shannon-McMillan-Breiman theorem allows to make a formal bridge between Boltzmann entropy and Shannon entropy rate: the number of typical sequences of length N behaves as $\mathcal{N}_N \sim 2^{Nh}$. For N large enough, this asymptotic relation writes $h \sim (1/N)\log_2 \mathcal{N}_N$, reminiscent of the Boltzmann entropy per particle $S_B/N = (k_B/N)\ln\Gamma_N$.

Boltzmann entropy S_B, defined at the level of N-particles microscopic configurations (phase space $\mathcal{X}^N$), should not be confused with the Shannon entropy of the empirical distribution (nor-

malized histogram) of the individual states in $\mathcal{X}$ (the type $L_{\bar{x}}$ of the configuration $\bar{x}$, see § 3.1. The former is an entropy in the phase space $\mathcal{X}^N$, the latter is the entropy of a distribution $L_{\bar{x}}$ in the individual state space $\mathcal{X}$, namely $H(L_{\bar{x}}) = -\sum_x (n_x/N)\log(n_x/N)$ where n_x is the number of particles (among N) in the individual state x (Mugur-Schächter 1980; Georgii 2003; Cover and Thomas 2006).

Thermodynamic entropy is derived (actually, *postulated*) from phenomenological considerations based on the (observed) second principle of the thermodynamics (Clausius 1865; Gallavotti 2006). A major achievement of Boltzmann is the identification of Boltzmann entropy with thermodynamic entropy: Boltzman entropy of the macrostate (U, V, N) coincides at the leading order (in the thermodynamic limit $N \to \infty$) with the thermodynamic entropy $S_{th}(U, V, N)$, providing a microscopic interpretation to Clausius entropy. It is justified by comparison of the macroscopic predictions of statistical mechanics with the empirical laws of thermodynamics, e.g. the expression of the entropy of the ideal gas. The identification requires the multiplicative factor k_B (equal to the ideal gas constant divided by the Avogadro number) in the definition of S_B, compared to a dimensionless statistical entropy. A *definition* of physical entropy is only possible in the framework of quantum mechanics, exploiting its intrinsically discrete formulation. Introducing the density matrix $\widehat{D}$ characterizing the quantum state of the system (a positive Hermitian operator with trace 1), entropy is defined as $S(\widehat{D}) = -k_B \mathrm{Tr}(\widehat{D}\ln\widehat{D})$. This entropy, introduced by Von Neumann in 1927, measures our ignorance about the system, and accordingly vanishes in case of a pure state (described by a single wave function) (Wehrl 1978; Balian 2004; Balian 2005). It is moreover a absolute entropy which vanishes at zero temperature, in agreement with the Nernst principle. It is another matter to assess whether it is useful a physical quantity, and how it can be measured and related to the macroscopic (thermodynamic) entropy. Another interpretation of this entropy is to measure the amount of information gained in a quantum measurement (yielding a pure state). When considering an evolving system ($\imath\hbar\, d\widehat{D}_t/dt = [\widehat{H}, \widehat{D}_t]$, $S(t) = \mathrm{Tr}(\widehat{D}_t \ln \widehat{D}_t)$ remains constant. Any reduction to essential variables of the density operator yields a reduced operator $\widehat{D}^0_t$, for which the associated entropy $S(\widehat{D}^0_t)$ increases. We refer to (Wehrl 1978; Balian 2004; Balian 2005) for a detailed discussion of the quantum-mechanical entropy.

8.3. *Maximization of Boltzmann entropy and large deviations*

Any macroscopic variable m appears as an additional constraint in defining the microcanonical ensemble. A Boltzmann entropy $S(m, U)$ can be associated to this reduced ensemble. To each value of m corresponds a "shell" of volume $e^{S(m,U)/k_B}$ in the complete microcanonical space (for the energy value U). The distribution of m is thus given by the large deviation formula, already derived by Einstein in 1925 (Einstein 1910):

$$P(m|U) = e^{\Delta S(m,U)/k_B} \tag{75}$$

where $\Delta S(m, U) = S(U, m) - S(U) \leq 0$ is proportional to the number N of particles, i.e. to the size of the system. In the limit as $N \to \infty$, the macrostate distribution becomes sharply peaked around the value m^* giving the maximum Boltzmann entropy. This property is essentially follows from a concentration theorem. It reflects the fact that, in the thermodynamic limit $N \to \infty$, an exponentially dominant fraction of microscopic configurations are associated to the macroscopic variable m^*. For N large enough, the distribution is sharply peaked and the typical behavior can be identified with the most probable behavior. In other words, one observes the most probable macrostate. At leading order in N, m^* is also the average value of the macrostate m. Note that (75) is a large deviation formula, with $\Delta S(m, U)/k_B$ as a large deviation function (Ellis 1985; Touchette 2009): it is not restricted to values of m close to m^*.

Arguments based on Boltzmann entropy explain the *irreversibility* of the relaxation of an isolated

system from a prepared state towards an equilibrium state, e.g. the fact that your coffee always cools and never gets heat from the surroundings, despite the invariance upon time reversal of the microscopic dynamics. (Lebowitz 1993a; Castiglione *et al.* 2008; Schulman 2010). Liouville theorem indeed ensures the constancy in time of the density in the microcopic phase space. The answer has been yet given by Boltzmann (Boltzmann 1877). Basically, the origin of this irreversibility lies in the non typicality of the initial configuration when considered in the final conditions, whereas the final equilibrium state is typical. This asymmetry is quantified by means of Boltzmann entropy of the two macrostates, amounting to compare the volumes of the phase space regions Γ_i and Γ_f associated respectively with the prepared initial state and the final equilibrium state (Lebowitz 1993a). Trajectories starting in Γ_i mostly evolve to Γ_f Time-reversed trajectories starting in Γ_i also mostly evolve to Γ_f In both cases, the odds for evolving to Γ_i rather than Γ_f is

$$\frac{|\Gamma_i|}{|\Gamma_f|} = e^{-(S_B^f - S_B^i)/k_B} \tag{76}$$

Accordingly, the spontaneous evolution corresponds to increasing Boltzmann entropy, and the probability of the time-reversed evolution (i.e. starting in Γ_f and evolving to Γ_i) is exponentially small in the thermodynamic limit $N \to \infty$. One can find in the literature statements like: "obviously, the outcome cannot carry more information hence its entropy cannot be smaller than the initial one", taken as an explanation of the observed irreversibility. This argument is misleading since information is not a conserved quantity, but rather a relative and context dependent notion. Here he information about the initial state refers to the missing information with respect to the knowledge of the initial context and similarly, information about the final state refers to the knowledge of the (different) final context and constraints.

8.4. *Boltzman-Gibbs entropy and the canonical ensemble*

In statistical mechanics textbooks (Chandler 1987), the canonical ensemble is currently derived by imposing a fixed average energy on otherwise equiprobable microscopic configurations. Jaynes (Jaynes 1957) underlined yet long ago taht statistical mechanics can also be derived from the maximum entropy principle, within a purely information-theoretic framework.. As presented in a general setting in § 3.4, this principle allows to determine the least biased distribution satisfying a given set of constraints on distribution moments. When applied to the velocity distribution of N independent and identical particles at fixed thermal energy (fixed mean square velocity, vanishing mean velocity), it yields the acknowledged *Maxwelll velocity distribution*:

$$\rho_N(\bar{v}_N) d^{3N}\bar{v}_N = \prod_{i=1}^{N} \rho_1(\vec{v}_i) d^3\vec{v}_i \quad \text{where} \quad \rho_1(\vec{v}) d^3\vec{v} = e^{-mv^2/k_BT} \left(\frac{m}{2\pi k_B T}\right)^{3/2} dv_x dv_y dv_z \tag{77}$$

(where v^2 is the square modulus of $\vec{v}$, namely $v_x^2 + v_y^2 + v_z^2$ in Cartesian coordinates), in agreement with the expression defining thermal velocity :

$$\langle mv^2/2 \rangle = 3k_BT/2 \tag{78}$$

When applied to configurations $\bar{x}_N$ and internal energy $E(\bar{x}_N)$, entropy maximization principle yields the well-known *Boltzmann-Gibbs distribution* in the microscopic phase space $\mathcal{X}^N$:

$$P(\bar{x}_N|\beta) = \frac{e^{-\beta E(\bar{x}_N)}}{Z(N,\beta)} \quad \text{with} \quad Z(N,\beta) = \int_{\mathcal{X}^N} e^{-\beta E(\bar{x}_N)}\, d\bar{x}_N \tag{79}$$

where $d\bar{x}_N$ is the element of integration in the $6N$-dimensional phase space (microscopic configuration $\bar{x}_N$ of N particles). Note the factorization of the distributions for the velocity and position degrees of freedom (respectively Maxwell and Boltzmann-Gibbs distributions), allowing to decouple kinetic theory and equilibrium statistical mechanics. Compared to the microcanonical ensemble, Boltzmann-Gibbs distribution gives different weight to the microstates, defining the

canonical ensemble gives different weight to the microstates. Nevertheless, their predictions for the thermodynamic quantities coincide in the thermodynamic limit $N \to \infty$ (Chandler 1987).

At a mesoscopic level, it is no longer relevant to describe the distribution of the microscopic configurations. Partial integration in energy shells $dE = \sum x, E(x) \in [E, E+dE]$ involves the microcanonical weight $e^{S_B(E,N)/k_N}$ of each shell, and it yields the distribution:

$$p(E \mid N, \beta) = \frac{e^{-\beta E}\ e^{S_B(E,N)/k_N}}{Z(N,\beta)} \tag{80}$$

Steepest-descent approximation of the partition function in the thermodynamic limit $N \to \infty$,

$$Z(N,\beta) = \int e^{-\beta E}\ e^{S_B(E,N)/k_N} dE \tag{81}$$

exploiting the extensivity of S_B (Touchette 2009), demonstrates that the dominant contribution is given by the maximum E^*, which also coincides with the average energy $\langle E \rangle \equiv U$ in the limit $N \to \infty$. Consistency with classical thermodynamics leads to identify $F = -(1/\beta) \ln Z(\beta, N)$ with the *free energy*, multiplier β with inverse temperature $1/k_B T$ and Boltzmann entropy at the maximum $E^* \equiv U$ with thermodynamic entropy, according to the relation $F = U - TS_{th}$.

Maximum entropy principle could also be applied to the inference of the distribution of energy levels at fixed average energy. It yields:

$$p_i = \frac{e^{-\beta E_i}}{Z(\beta)} \tag{82}$$

with a caveat: the energy levels have to be discrete and non degenerate. Indeed, the application of maximum entropy principle at fixed average energy $\langle E \rangle$ to a continuous energy density $p(E)dE$ yields an inconsistent result: it misses the density of states. We here recover that maximum entropy principle and more generally Shannon entropy are well-defined and can be used safely only in discrete spaces of states (§ 3.4). As mentioned above, § 8.2, the only rigorous foundation of statistical entropy lies at the quantum level, and other notions are derived by coarse-graining and projections in a more or less approximated way (Wehrl 1978; Balian 2004).

Another notion of entropy is encountered in statistical mechanics, namely *Gibbs entropy*. It is defined as:

$$S_G(t) = \int \rho(\bar{x}_N, t) \ln \rho(\bar{x}_N, t) d^{6N}\bar{x}_N \tag{83}$$

where $\bar{x}_N$ is the system position in the full microscopic phase space (configuration with $6N$ degrees of freedom, for both the positions and the velocities of the N particles of the system), and $\rho(\bar{x}_N, t)$ the density describing the probability of presence of the system in this phase space. It has nevertheless two flaws: it is defined up to an additive constant (as mentionned in § 2.6, the continuous extension of Shannon entropy is not invariant with respect to a cooordinate change) and Liouville theorem for the microscopic dynamics ensures that $S_G(t)$ remains constant in time, even in conditions where the Second Principle of the thermodynamics predicts an increase of the thermodynamic entropy. Accordingly, Gibbs entropy in this form cannot be identified with thermodynamic entropy. Both flaws are cured by considering (Castiglione *et al.* 2008) a coarse-grained version of Gibbs entropy, that is, Shannon entropy of a distribution describing the location in the microscopic phase space with a finite resolution. It can be shown that this coarse-grained version increases in time with a rate related to the metric entropy, § 7.1, of the microscopic dynamics. We refer to (Castiglione *et al.* 2008) for a detailed discussion (deserving several chapters, far beyonf the scope of the present review) of the connections between statistical mechanics and chaos theory.

8.5. *Dissipative structures and minimum entropy production principle*

Prigogine (Prigogine 1967; Nicolis and Prigogine 1977) developed the notion of *dissipative structure* (although examples were evidenced and studied well before his work, e.g. Bénard cell). It refers to an organized pattern arising in open systems, in which local order appears at the expense of energy or matter input. Thermodynamic entropy S_{th} is defined only for equilibrium states. Out of equilibrium, one could define an *entropy production rate*. The entropy production decomposes according to $dS_{th} = dS_{irr} + dS_{exch}$ where dS_{exch} is the contribution following from exchanges of matter and energy. At steady state $dS_{th} = 0$, but it is possible to have $dS_{irr} > 0$ at the expense of $dS_{exch} < 0$, which is precisely the case of dissipative structures. dS_{irr} is often seen as a measure of the irreversibility of the system. Its definition and interpretation are nevertheless restricted to the framework of thermodynamics of irreversible processes, embedded in linear response theory.

Within this framework, Prigogine introduced a *minimum entropy production principle* (not to be confused with the maximum entropy principle for statistical inference): $(d/dt)[(dS_{th}/dt)_{irr}] = 0$ where $(dS_{th}/dt)_{irr}$ is the entropy production rate due to irreversible processes (Prigogine 1967). Nevertheless, this principle, expressing the stability of nonequilibrium steady state, is rigorously derived only under very restrictive conditions (assumptions of local equilibrum thermodynamics and linear response, isotropic medium, time independence of boundary conditions and linear response coefficients, isothermal system in mechanical and thermal equilibrium). Its general validity and application are thus highly questionable (Kay 1984). As emphasized in (Mahara and Yamaguchi 2010), while entropy production could be used to discriminate different patterns, minimizing entropy production is not a valid criterion of pattern selection.

We refer the reader to the very deep and stimulating analysis proposed by Jaynes thirty years ago (Jaynes 1980). He pointed out that Kirchhoff laws for determining the distribution of currents in an electric circuits are already fully determined by conservation laws, with no need of an additional entropic criterion. He asked the question of the nonequilibrium extension of Gibbs work on the characterization of heterogeneous equilibrium (phase coexistence) using a variational principle on thermodynamic entropy. It is to be underlined that at this point, all the derivations and justifications of minimum entropy principle (by Onsager, Prigogine and followers) are based on linear response theory, where the evolution is ruled by linear relations between fluxes and forces.

8.6. *Nonequilibrium systems and chaotic hypothesis*

A general definition of entropy and entropy production in far-from-equilibrium system, beyond linear response theory, requires to start at the more basic level of the microscopic dynamics (Ruelle 2003; Gruber *et al.* 2004). Within such a dynamic view of irreversible processes, it is currently assumed that the dynamics is well described by an hyperbolic dynamical system (Cohen and Gallavotti 1999; Evans and Searles 2002; Gallavotti 2006). This so-called *chaotic hypothesis* is the far-from-equilibrum analog of the assumption of ergodicity or molecular chaos (assumption of microscopic decorrelation). The local rate of entropy production $e(x)$ is then equal to the local rate of phase space volume contraction at point x. The global rate of entropy production is obtained by integrating $e(x)$ over the whole phase space, according to the weight given by the nonequilibrium steady state measure $\rho(dx)$, namely $\int e(x)\rho(dx)$ (Ruelle 2003).

Gaspard introduced the Shannon time-reversed entropy rate (Gaspard 2004):

$$h^R = \lim_{n\to\infty} (1/n) H_n^R \tag{84}$$

with $H_n^R = -\sum_{\bar{w}} p_n(\bar{w}) \log_2 p_n(\bar{w}^R)$, $\bar{w} = (w_1, \ldots, w_n)$and $\bar{w}^R = (w_n, \ldots, w_1)$. Then he shows

that for Markov processes (at least), the entropy production rate writes:

$$dS/dt = h^R - h \quad \text{where} \quad S(t) = -\sum_w p_t(w) \log_2 p_t(w) \tag{85}$$

This time-reversal symmetry breaking, reflecting in entropy production, corresponds to the fact that the distribution of incoming and ougoing particles strongly differ. The latter is finely correlated due to the interactions between the particles inside the system. Observing the time-reversed steady-state would require to prepare the incoming flux of particles according to such intricately correlated distribution. This formula (85) relates in a rigorous and quantitative way irreversibility and entropy production rate in the considered nonequilibrium stationary state. We underline that the irreversibility of a system driven far-from-equilibrium by fluxes s fundamentally different from the irreversibility observed in the relaxation of an isolated system after lifting a constraint discussed in § 8.3.

8.7. *Thermodynamic cost of computation*

Beyond the formal link based on maximum entropy inference of the Boltzmann-Gibbs distribution (Jaynes 1957), another relationship between statistical mechanics and information theory is the paradox termed Maxwell's demon. It was first pointed out by Szilard (Szilard 1929). N particles are evenly distributed in two compartments of the same size but initially at different temperatures $T_1 < T_2$. The demon stands in the hot compartment near a door between the two compartments. He lets in particles from the cold compartment if their velocity is higher than $\sqrt{3k_BT_2/m}$, and lets out particles of the hot compartment if their velocity is higher than $\sqrt{3k_BT_1/m}$. In this way the hot compartment gets hotter and the cold compartment gets colder, against the prescription of the second principle. Brillouin (Brilllouin 1951a; Brillouin 1951b) suggested a way out Maxwell's demon paradox by showing, in a specific example, that work has to be performed in order to achieve a measurement. Put in other words, the demon needs information about the particle velocity, which has a cost (Brillouin 1956). In a simpler variant, compartments are at the same temperature. The demon only lets particles in, so that in the final state of the system, all N particles in only one compartment. The decrease of thermodynamic entropy by $k_BN \ln 2$, equal to the amount of information required to know the position of each particle. In a measurement, entropy increases by an amount at least equal to the information gained (Balian 2004).

Later, Landauer (Landauer 1961) proposed another solution of the puzzle, giving a lower bound on the work required for memory erasure. Zurek then showed that algorithmic complexity sets limits on the thermodynamic cost of computation (Zurek 1984). Recently, Sagawa and Ueda (Sagawa and Ueda 2009) unified these different results by demonstrating a general inequality: $W_{meas} + W_{eras} \geq k_BT\ I$ where W_{meas} is the work cost of the measure, W_{eras} that of erasing the memory storing the result, and I the mutual information between the measured system and the memory (i.e., the information gained on the system in the measurement). The work W_{meas} could vanish in some instances, in which case Landauer's result is recovered, while the whole inequality is also consistent with Brillouin's result.

9. Typicality and statistical laws of collective behavior

9.1. *Probabilistic modeling and subjective probabilities*

The notion of statistical entropy, being relative to a probability distribution, leads to question the very foundations of probability theory and the epistemic status of a probabilistic description (Mugur-Schächter 1980). In practice, the question can be focused on the reconstruction in a given experimental situation of the relevant probability distribution.

A well-known alternative the reconstruction and epistemic status of a probability distribution is the alternative between *frequentist* and *subjective* (or Bayesian) viewpoints (Jaynes 1957; Bricmont 1995). Both viewpoints yield efficient reconstruction methods. The frequentist viewpoint is the realm of statistical estimation from independent samples, essentially based on the law of large numbers (Samengo 2002). The Bayesian viewpoint is the realm of learning and recursive algorithms, updating with data a prior distribution into a posterior one. A seminal paper by Cox (Cox 1946) underlined that the frequentist definition is indissociable with the existence of an ensemble (at least conceptually). The Bayesian viewpoint is termed there the idea of "reasonable expectation". It is related to the notion of "degree of rational belief" formulated by Keynes. Some Bayesian probabilities cannot be cast in an ensemble (i.e. frequentist) viewpoint. Cox cited the inspiring examples of the probability that there exists more than one solar system; the probability that a physical constant lies within some bounds (today formulated as an "estimation problem"); and the probability that some property in number theory is true when considering all integers. The non scientist will rather think of the probabilistic proof of the existence of Santa Klaus featured by Howard Buten (Buten 1989).

Jaynes (Jaynes 1973; Jaynes 1982b) already underlined the alternative between the frequentist view, trying to estimate the frequencies of the various events, and the subjective view, aiming at determining the probability distribution that describes our state of knowledge. In this regard, information theory provides a constructive criterion, the maximum entropy principle (Jaynes 1982a), for setting up probability distributions on the basis of partial knowledge, discussed in (Jaynes 1957) in the context of statistical physics.

The subjective view on probabilities (Cox 1946; de Finetti 1970; Gillies 2000; Balian 2005) encapsulates a priori but incomplete knowledge, e.g. a set of possible states, but also apparent randomness at the observation scales. In both cases, it means that our limited perception is best represented by a probability distribution notwithstanding whether the nature of the system is stochastic or not. The probabilistic aspect of the description is only that of our representation of the reality, with no ambition of saying something about the nature of the real phenomenon. A probability distribution does not aim at being an intrinsic and absolute character ruling the system behavior (as it is in quantum mechanics), but only the most operational and faithful account of *our* knowledge on the system. We refer to (Jaynes 1957; Jaynes 1973) for a detailed and substantiated discussion of this viewpoint. Such a pragmatic view on probabilistic description is currently adopted for chaotic dynamic systems (Nicolis and Gaspard 1994) when one gives up a description in terms of deterministic trajectories for a stationary and global description in terms of invariant measure (the latter is nothing but the distribution describing the probability of presence in the phase space). In any case, probability theory can be seen as a mere operational tool, even is there is no stochasticity involved in the problem, as in the probabilistic formulation of some properties in number theory (Cox 1946; Ford 2007). Interpretation of entropy is thus far more natural in the subjective viewpoint where p describes our partial knowledge (and partial ignorance) of the outcome. Entropy then measures the uncertainty of the observers.

It is to note that in the realm of *classical* physics, we cannot assess whether a system is intrinsically probabilistic, except in starting at the quantum level (Lesne 2007). But the observed randomness, say, of a coin tossing, is of very different nature than quantum uncertainty. Mostly, it can be accounted for by arguing of the chaotic nature of the coin motion while flipping in the air. The randomness thus originates in our lack of knowledge of the initial conditions and countless minute influences experienced by the coin during its tossing. The probabilistic nature is not that of the system but that of one of our possible descriptions. In particular, it depends essentially on the scale of the description. An acknowledged example is diffusion for which a hierarchy of descriptions exist, according to the scale and the level of coarse-graining, starting from a deterministic reversible

description (molecular dynamics) to several stochastic descriptions (master equation, random walk and Fokker-Planck equation, Langevin equation) to a deterministic irreversible description (Fick law and the century-old diffusion equation) (Castiglione *et al.* 2008).

Finally, let us think to a binary event, described by a Boolean variable X. The statistical features of this variable are fully captured by a single real number $p \in [0,1]$ describing the probability that $X = 1$. In the case of a structured population, distinguishing explicitly subpopulations α with fraction f_α (hence $\sum_\alpha f_\alpha = 1$) allows to describe some heterogeneities in the process yielding the value of X, by considering a specific value p_α in each sub-population. We are thus faced to an alternative between a detailed description by means of an array $[(p_\alpha)_\alpha$ and a global probabilistic view, namely an effective description of the knowledge available at the scale of the population involving a single number $p = \sum_\alpha p_\alpha f_\alpha$. This effective quantity p describes the probability that an individual chosen at random in the overall population takes the value $X = 1$, while p_α describes the probability that an individual chosen at random in the subpopulation α takes the value $X = 1$. This abstract example illustrates the existence of nested probabilistic descriptions, preventing any further attempt of a would-be "intrisic stochastic nature" of a system. We deal only with models, i.e. abstractions and representations of the reality. Our statements thus refer to models, and are pertinent to the reality only insofar as it is properly captured by the model.

9.2. *Statistical laws and collective behaviors in physics*

We have just argued that probability, in the subjective viewpoint, is a privileged framework allowing to account in a unified way of observation scales and the limits they set on our perceptions and representations. It is also an unified framework to investigate collective behaviors and unravel the mathematical structures underlying emergent properties. A central physical example is thermodynamic behavior. It corresponds to a sharp distribution for macroscopic quantities, meaning that almost all microscopic configurations yields the same macroscopic values. In such a case, the probability distribution of the microscopic configurations (i.e. their respective frequencies of occurrence) has almost no macroscopic consequence, as soon as it is non singular. Accordingly, thermodynamics relies almost uniquely on universal statistical laws, mainly the law of large numbers and the central limit theorem. The counterpart of this universality is that macroscopic behavior is quite insensitive to microscopic features. In particular, knowing the macroscopic behavior gives no insight on the microscopic distribution and is useless to infer some knowledge about the microscopic elements. The success of the maximum entropy approach evidences that thermodynamical laws originally rely on universal statistical laws ruling the structure and features of emergent behaviors, rather than from specific physical laws (Jaynes 1957).

More generally, statistical laws express rules of collective behavior, no matter the physical nature of the elements and their interactions. They state a general *mathematical* property of *any* high-dimensional system (e.g. many-body systems in physics or long messages in communication theory). They account for instance for the ubiquitousness of Guassian distributions (resulting from the central limit theorem). The same all-or-none law arises in different contexts and under different names (§ 3.1, § 5.1):
— law of large numbers and Lévy *all-or-none law* (Lévy 1965) in probability and statistics;
— *concentration theorems* (Robert 1990) in probability but also in geometry and functional analysis (Gorban 2007);
— *asymptotic equipartition property* (Shannon 1948; Cover and Thomas 2006) in information theory;
— *ergodic theorem* in dynamical systems.
Close connections can be established between these different laws (Lesne 1998). They can be seen as an universal mathematical structure of collective behaviors.

Let us consider again Shannon-McMillan-Breiman theorem, § 5.1. The property that $\lim_{n\to\infty}(1/n)\log_2 \hat{P}_n - h = 0$ is an asymptotic property, insofar as modifying a finite number of random variables does not change whether it is true or false; in particular it is exchangeable, meaning that it is unaffected by any permutation of a finite number of terms. Shannon-McMillan-Breiman theorem, when restricted to a stationary uncorrelated source, is thus an instance of the *all-or-none law* established by P. Lévy (Lévy 1965), also known as the *Hewitt-Savage 0-1 law*. It states that an asymptotic property of a sequence of independent and identically distributed random variables is true with probability either 0 or 1. Here $\lim_{n\to\infty}(1/n)\log_2 \hat{P}_n(\bar{x}) - h = 0$ is true with probability 1 whereas for any $h' \neq h$, $\lim_{n\to\infty}(1/n)\log_2 \hat{P}_n(\bar{x}) - h' = 0$ has a null probability to be true.

Predictability and simplicity of macroscopic physical phenomena come from the fact that at the macroscopic level, a wealth of behaviors result from a bottom-up integration and emergence. They are ruled by simple statistical laws and a simple description is available. Macroscopic properties are then almost fully defined by statistical laws and geometrical constraints. Physics is involved only in prescribing the *universality class* of the emergent behavior. Basically, one has to discriminate between systems with short-range correlations, displaying scale separation between microscopic and macroscopic levels, and systems with long-range correlations, associated with criticality and anomalous statistical laws (Lesne 1998; Castiglione *et al.* 2008). A typical example is *diffusion*, passing from normal to anomalous in case of long-range correlations (self-avoiding walks), which corresponds to the passage from the universality class of the Wiener process to that of fractal Brownian motions. Another anomaly is observed in diffusive behavior when the variance of the elementary steps diverge, corresponding to the passage from the central limit theorem assessing convergence to a Gaussian distribution to generalized limit theorems assessing convergence to Lévy stable laws (Lesne 1998; Castiglione *et al.* 2008). In general universality and robustness arise in physics as soon as statistics and geometry aresufficient to determined the emergent features. A typical exemple is provided by percolation lattices (Lesne 1998; Laguës and Lesne 2008). Microscopic details only matter insofar as they control the universality class to which the system belongs.

9.3. *Typicality*

Several notions of typicality can be considered, some of which have been already encountered in the previous paragraphs.

1) A notion based on concentration theorems, § 3.1, for a configuration or a sequence $(X_i)_i$ of independent and identical elements. Reasoning on the configuration or sequence type, typical sequences belong to the type that is the most populated. Conversely, sequences are exceptional (non typical) when their type is represented by a vanishing fraction (exponentially small as a function of the number of elements in the configuration or the length of the sequence) compared to the most populated one. The law of large numbers can be seen as a statement about the typical behavior of the empirical average $\widehat{m}_N = (1/N)\sum_i^N X_i$. Namely for any arbitrary small $\epsilon > 0$ and $\delta > 0$, there exists $N_{\epsilon,\delta}$ such that for $N > N_{\epsilon,\delta}$, the probability of the realizations of the sequence satisfying $|\widehat{m}_N - m] < \epsilon$ is smaller than δ, meaning that asymptotically, almost all realizations of the sequence are typical as regards the behavior of the empirical average.

2) A notion based on Sanov theorem, § 3.2, for sequences of independent and identical elements: a pair of sequences $(\bar{x}_N, \bar{y}_N)$ is *jointly typical* if each individual sequence is typical, respectively with respect to h_X and h_Y and if $\mid -(1/N)\log_2 P_N(\bar{x}_N, \bar{y}_N) - h_{X,Y} \mid$ is small. Given a joint distribution $p(x, y)$, the probability that a pair of independent and identically distributed sequences $(\bar{x}_N, \bar{y}_N)$ drawn according to the product distribution $q(x, y) = p(x)p(y)$ seems to be typical with respect to the joint distribution p is asymptotically equivalent to $2^{-ND(p||q)} = 2^{-NI(X,Y)}$.

3) A notion based on a generalized asymptotic equipartition property, namely the fact that almost

surely $\lim_{N\to\infty}(1/N)\log_2[p_0(\bar{x}_N)/p_1(\bar{x}_N)] = D(p_0||p_1)$. Hence a sequence $\bar{x}_N$ of length N is said to be *relative-entropy typical* if $(1/N)\log_2[p_0(\bar{x}_N)/p_1(\bar{x}_N)]$ is close to $D(p_0||p_1)$.

4) A notion based on Shannon-McMillan theorem, § 5.1, for correlated sequences. A sequence $\bar{x}_N$ of length N generated by a source of entropy rate h is typical if $\mid -(1/N)\log_2 P_N(\bar{x}_N) - h \mid$ is small. The realizations whose probability satisfies Shannon-McMillan-Breiman estimate form the typical set (strictly, defined once some tolerance ϵ is given), quite small but of probability close to 1. For correlated binary sequences of length N, one has 2^N possible realizations but only about 2^{Nh} typical ones.

5) The notion of typicality also arises in connection to the *ergodic theorem of Birkhoff*. Let us recall that a triplet $(\mathcal{X}, f, \mu)$ composed of a transformation f on the phase space $\mathcal{X}$ with invariant measure μ is ergodic is any f-invariant subset is of either full or null measure. Then for any functional ϕ from $\mathcal{X}$ to $\mathbf{R}$, there exists a subset $\mathcal{X}_\phi$ of full measure (that is, $\mu(\mathcal{X} - \mathcal{X}_\phi = 0)$ such that for any $x \in \mathcal{X}_\phi$, $\lim_{N\to\infty}(1/N)\sum_{i=0}^{N-1}\phi[f^i(x)] = \int_{\mathcal{X}}\phi(x)d\mu(x)$. In this sense, the elements of $\mathcal{X}_\phi$ are typical since their behaviors are all identical and coincide with an average quantity, namely time averages along a typical trajectory equals ensemble averages. We have encountered in § 6.2 another ergodic theorem (Ziv and Lempel 1978), endowing typical sequences with an additional property: For a stationary ergodic finite-state source, almost all sequences share the same algorithmic complexity (hence the same randomness) which coincides with the entropy rate of the source, up to a normalization factor.

The converse of typicality is rarity. Exceptional events are non typical events. Several notions nevertheless overlap and should be carefully distinguished. The sequence 123456789 can be termed exceptional because it is of low complexity, namely a short program is able to generate it. It is also intuitively atypical, or exceptional, insofar as one implicitly compare the number of sequences $(n, n+1, \ldots, n+8)$ to the set of sequences that are not of this form. In other words, one compares the types of the sequences rather than the sequences themselves. This yields two ways of being random: either having the largest algorithmic complexity, either belonging to the type the most represented. These two viewpoints are in fact related: only an asymptotically vanishing fraction of sequences of length $N \to \infty$ can be generated by a program shorter than the typical length Nh equal to the length of the programs generating typical sequence. This is exactly the meaning of algorithmic complexity. In this viewpoint, typicality coincides with (full) randomness. Note that in all cases typicality is an asymptotic feature, well-defined only in the limit $N \to \infty$ where N is the sequence length or number of elements. Arguments of genericity and typicality are ubiquitous in statistical physics, but we expect that they cannot be applied blindly in biology, where rare events could play essential role, see § 9.5.

9.4. *Entropy, order and disorder*

Let us discuss in what respect entropy can be seen as a *measure of disorder*. This current and appealing statement is indeed flawed with some fuzziness, since it requires to first define what is meant by disorder, beyond the plain and non technical meaning of the word. Two viewpoints can be envisioned, recovering the alternative between the (statistical) information-theoretic approach and the algorithmic one. A first view is that order (e.g., for a configuration of N elements) is associated with the existence of a simple generating rule (Dessallles 2006). For instance, the sequence 123456 is ordered insofar as it is generated by the simple rule $x_{n+1} = 1 + x_n$. The presence of a structure or pattern (in space or time) reflects a symmetry breaking with respect to the full symmetry of an homogeneous/stationary distribution invariant with respect to any translation. Specifying a structure amounts to specify a lack of invariance. This corresponds to a decrease of entropy compared to the fully random case (Leyton 2001). Another view is that speaking of order and disorder amounts to compare sets. For instance, the sequence 123456 is ordered insofar as it is a

representative of the set $\{(n, n+1, \ldots, n+5)\}$, opposed to its complement (any sequence that is not of the form $(n, n+1, \ldots, n+5)$). The difficulty with this view is to require a prior and necessarily subjective delineation of an ensemble from a single observation. Order and disorder are then relative to the mind who perceives them.

Gorban (Gorban 2007) gives the very inspiring example of a castle, a garden of stones and any pile of stones: the castle is to be compared to any pile of stones that is not a castle, but for the gardener, the garden of stones has also to be compared to any pile of stones that is not that garden of stones. As such, a garden of stones is as ordered and non typical than a castle, whereas it is less ordered when using the criterion of individual configuration complexity. A garden of stones is seen very differently by the gardener and a foreigner. In a formalized way, the relevant entropy is that of the coarse-graine distribution associated with a partition of the space into weighted subsets (currently the weight is simply the cardinal). Disorder appears as a lack of specific features, structures, patterns, so that the class of configurations looking like the given one is very large; the given configuration is then termed disordered. Rather than a measure of disorder, entropy is a measure of the typicality of the disorder, i.e. a measure of degeneracy: how many configurations share the same macroscopic observables and constraints) or a measure of degeneracy.

9.5. *Beyond physics ... application to living systems ?*

It is now acknowledged that the basic role of food is to provide enough energy to the organism to free itself from the entropy produced while it is alive. In this respect, a living organism is an instance of dissipative structure, see § 8.5. It is essential that the balance is written in terms of free energy. Boltzmann already pointed out that life is a struggle for entropy [§]. His view was expanded on by Schrödinger with the concept of negative entropy (the opposite of an entropy) (Schrödinger 1944), of *negentropy*, a term dubbed by Brillouin (Brillouin 1953). The problem with such formulations is that entropy of a driven system (open system driven far from equilibrium by fluxes) is not defined (Ruelle 2003), and the second principle, on which Schrodinger's statement implicitly refers to, does not directly apply to open systems, in particular living systems. This statement has thus to be taken as an intuitive understanding but not as a technical nor constructive theory.

Another caveat concerns the use of maximum entropy methods: they rely on a genericity argument: the empirically observed configuration is one of the typical ones, hence it is legitimate to identify its type (empirical distribution) with p^* maximizing $H(p)$. But a genericity argument, currently valid for physical systems, is highly questionable for living systems, whose behavior has been fine-tuned by biological evolution into very specific regimes, involving the non generic co-adaptation of several parameters.

Finally, universal statistical laws, § 9.2, underlying thermodynamic behavior are valid in physical systems under the quite mild condition that correlations between elements are summable. They fail only at critical points and are then replaced by self-similar features (Laguës and Lesne 2008; Castiglione *et al.* 2008). By constrast, their validity is highly questionable in complex systems, in particular living systems, due to top-down causation. We here mean the existence of feedback loops by which collective behaviors and emergent features influence back not only the elementary states, but also their rules of interaction and evolution. Such feedbacks from the macroscopic level to underlying levels prevent the law of large numbers and central limit theorem to apply. At the present day, information theory is only bottom-up, and is neither suited to take into account how

[§] The specific quote is (Boltzmann 1877; Cercignani 1988b): The general struggle for existence of animate beings is not a struggle for raw materials — these, for organisms, are air, water and soil, all abundantly available — nor for energy, which exists in plenty in any body in the form of heat, but of a struggle for entropy, which becomes available through the transition of energy from the hot sun to the cold earth.

an emerging feature modifies the state space and rules of an element. A first direction would be to change the level of the description and investigate relations between distribution of probabilities in order to capture invariants and predictable facts (Lesne and Benecke 2008). Another direction is to focus on interlevel relations and consistency, in the hope that some universality lies in the regulatory, schemes which is absent when restricting to a single level of organization. In any case, novel statistical laws, centrally involving the reciprocal coupling and consistency between the different levels of organization have to be developed.

References

Algoet, P. H. and Cover, T. M. (1988) A sandwich proof of the Shannon-McMillan-Breiman theorem. *Ann. Prob.* **16**, 899–909.

Amari, S. and Nagaoka, H. (2000) *Methods of information geometry*, Oxford University Press.

Avery, J. (2003) *Information theory and evolution*, World Scientific, Singapore.

Badii, R. and Politi, A. (1997) *Complexity. Hierarchical structures and scaling in physics*, Cambridge University Press.

Balding, D., Ferrari, P. A., Fraiman, R., and Sued, M. (2008) Limit theorems for sequences of random trees. *TEST* DOI 10.1007/s11749-008-0092-z. arXiv:math/0406280.

Balian, R. (2004) Entropy, a protean concept. In J. Dalibard, B. Duplantier and V. Rivasseau (editors), *Entropy*, Poincaré Seminar 2003, Birkhaüser, Basel, 119–144.

Balian, R. (2005) Information in statistical physics. *Studies in History and Philosophy of Modern Physics* **36**, 323–353.

Banavar, J. R., Maritan, A., and Volkov, I. (2010) Applications of the principle of maximum entropy: from physics to ecology. *J. Phys. Cond. Matt.* **22**, 063101.

Blanc, J. L., Pezard, L., and Lesne, A. (2011) Mutual information rate of pair of symbolic sequences. Submitted.

Blanc, J. L., Schmidt, N., Bonnier, L., Pezard, L., Lesne, A. (2008) Quantifying neural correlations using Lempel-Ziv complexity. In L. U. Perrinet and E. Daucé (editors), *Proceedings of the Second french conference on Computational Neuroscience (Neurocomp'08)*, ISBN 978-2-9532965-0-1, 40–43.

Boltzmann, L. (1877) ber die Beziehung zwisschen dem zweiten Haubtsatze der mechanischen Wrmetheorie und der Wahrscheinlichkeitsrechnung respektive den Stzen ber das Wrmegleichgewicht ("On the Relation between the Second Law of the Mechanical Theory of Heat and the Probability Calculus with respect to the Propositions about Heat-Equivalence"). *Wiener Berichte* **76** 373435. Included in *Wissenschaftliche Abhandlungen*, Vol. 2, paper 42 (1909) Barth, Leipzig; reissued in 1969, Chelsea, New -York.

Breiman, L. (1957) The individual ergodic theorem of information theory. *Ann. Math. Statist.* **28**, 809–811. Correction: **31** (1957) 809–810.

Bricmont, J. (1995) Science of chaos or chaos in science. *Physicalia Mag.* **17**, 159–208.

Brillouin, L. (1951) Maxwell's demon cannot operate: Information and entropy. *J. Appl. Phys.* **22**, 334–337.

Brillouin, L. (1951) Physical entropy and information. *J. Appl. Phys.* **22**, 338–343.

Brillouin, L. (1953) Negentropy principle of information. *J.Appl. Phys.* **24**, 1152–1163.

Brillouin, L. (1956) *Science and Information Theory*, Academic Press, New York.

Brin, M. and Katok, A. (1983) On local entropy. In J. Palis (editor), *Geometric dynamics*, Lecture Notes in Mathematics **1007**, Springer, Berlin, 30–38.

Brudno, A. A. (1983) Entropy and the complexity of the trajectory of a dynamical system. *Trans. Moscow Math. Soc.* **44**, 127–152.

Buten, H. (1989) *What to my wondering eyes*, Harper Collins, New York.

Callen, H. B. (1985) *Thermodynamics and thermostatics*, 2nd edition, Wiley, New York.

Castiglione, P., Falcioni, M., Lesne , A., and Vulpiani, A. (2008) *Chaos and coarse-graining in statistical mechanics*, Cambridge University Press.

Cercignani, C. (1988) *The Boltzmann equation and its applications*, Springer, Berlin.

Cercignani, C. (1998) *Ludwig Boltzmann — The man who trusted atoms*, Oxford University Press.

Chaitin, G.J. (1966) On the length of programs for computing finite binary sequences. *J. ACM* **13**, 547–569.

Chandler, D. (1987) *Introduction to modern statistical mechanics*, Oxford University Press.

Clausius, R. (1865) *The mechanical theory of heat — with its applications to the steam engine and to physical properties of bodies*, John van Voorst, London.

Cohen, E.G.D. and Gallavotti, G. (1999) Note on two theorems of nonequilibrium statistical mechanics. *J. Stat. Phys.* **96**, 1343–1349.

Cover, T.M. and Thomas, J.A. (2006) *Elements of information theory*, 2nd edition, Wiley, New York.

Cox, R.T. (1946) Probability, frequency, and reasonable expectation. *Am. J. Phys.* **14**, 1–13.

Csiszár, I. (1975) I-divergence geometry of probability distributions and minimization problems. *Ann. Prob.* **3**, 146–158.

Csiszár, I. (1998) The Method of types. *IEEE Trans. Inf. Th.* **44**, 2505–2523.

Csiszár, I. and Körner, J. (1981) *Information theory, coding theorems for discrete memoryless systems*, Akadémiai Kiadoó, Budapest.

de Finetti, B. (1970) *Theory of probability — a critical introduction treatment*, Wiley, Chichester.

Dessalles, J. L. (2006). A structural model of intuitive probability. In D. Fum, F. Del Missier and A. Stocco (editors), *Proceedings of the seventh International Conference on Cognitive Modeling*, Edizioni Goliardiche, Trieste, 86–91.

Durand, B. and Zvonkine, A. (2007) Kolmogorov complexity. In E. Charpentier, A. Lesne, and N. Nikolski (editors), *Kolmogorov's Heritage in Mathematics*, Springer, Berlin, 281–300.

Einstein, A. (1910) Ann. Phys. (Leipzig) **33** 1275–1298. English translation: "Theory of opalescence of homogeneous liquids and mixtures of liquids in the vicinity of the critical state. In J. Alexander (editor), *Colloid Chemistry*, Rheinhold, New York, 1913, Vol. I, 323–329. Reprinted in J. Stachel (editor), *The Collected Papers of Albert Einstein*, Princeton University Press, Princeton (1987), Vol. 3, 231–249.

Ellis, R. S. (1985) *Entropy, large deviations and statistical mechanics*, Springer, Berlin.

Evans, D. J. and Searles, D. J. (2002) The fluctuation theorem. *Adv. Phys* **51**, 1529–1585.

Falcioni, M., Loreto, V., and Vulpiani, A. (2003) Kolmogorov's legacy about entropy, chaos and complexity. In A. Vulpiani and R. Livi (editors), *The Kolmogorov Legacy in Physics*, Springer, Berlin, 85–108.

Feldman, D. P. (2002) A brief introduction to information theory, excess entropy and computational mechanics. Available online: http://hornacek.coa.edu/dave/

Feldman, D. P. and Crutchfield, J. P. (1998) Measures of statistical complexity: Why? *Phys. Lett. A* **238**, 244–252.

Ford, K. (2007) From Kolmogorov's theorem on empirical distribution to number theory. In E. Charpentier, A. Lesne and N. Nikolski (editors), *Kolmogorov's heritage in mathematics*, Springer, Berlin, 97–108.

Frank, S. A. (2009) The common patterns of nature. *J. Evol. Biol.* **22**, 1563-1585.

Gallavotti, G. (1998) Chaotic dynamics, fluctuations, nonequilibrium ensembles. *Chaos* **8**, 384–393.

Gallavotti, G. (2006) Entropy, thermostats and the chaotic hypothesis. *Chaos* **16**, 043114.

Gaspard, P. (2004) Time-reversed dynamical entropy and irreversibility in Markovian random processes. *J. Stat. Phys.* **117**, 599–615.

Gell-Mann, M. and Lloyd, S. (1996) Information measures, effective complexity, and total information. *Complexity* **2**, 44–52.

Gell-Mann, M. and Lloyd, S. (2003) Effective complexity. In M. Gell-Mann and C. Tsallis (editors), *Nonextensive Entropy — Interdisciplinary Applications*, Oxford University Press, 387–398.

Georgii, H.O. (2003) Probabilistic aspects of entropy. In A. Greven, G. Keller and G. Warnecke (editors), *Entropy* , Princeton University Press, 37–54.

Gillies, D. (2000) *Philosophical theories of probability*, Routledge, London.

Glasner, E. (2003) *Ergodic theory via joinings*, American Mathematical Society, Providence, § 15.8.

Gorban, A. N. (2007) Order-disorder separation: Geometric revision. *Physica A* **374**, 85–102.

Grassberger, P. (1986) Toward a quantitative theory of self-generated complexity. *Intl. J. Theor. Phys.* **25**, 907–938.

Gray, R. M. (1990) *Entropy and information theory*, Springer, New York. Available at

`http://ee.stanford. edu/~gray/it.html`

Gruber, C., Pache, S., and Lesne, A. (2004) On the Second Law of thermodynamics and the piston problem. *J. Stat. Phys.* **117**, 739–772.

Haegeman, B. and Etienne, R. S. (2010) Entropy maximization and the spatial distribution of species, *Am. Nat.* **175**, E74–E90.

Honerkamp, J. (1998) *Statistical physics*, Springer, Berlin, § 1.2.4.

Ihara, S. (1993) Information theory for continuous systems, World Scientific, Singapore.

Jaynes, E. T. (1957) Information theory and statistical mechanics Part I. *Phys. Rev.* **106**, 620–630. Part II, *Phys. Rev.* **108**, 171–190.

Jaynes, E. T. (1973) The well-posed problem. *Found. Phys.* **3**, 477–493.

Jaynes, E. T. (1979) Where do we stand on maximum entropy? In R. D. Levine and M. Tribus (editors), *The Maximum Entropy Formalism*,MIT Press, Cambridge MA, 15–118.

Jaynes, E. T. (1980) The minimum entropy production principle. *Ann. Rev. Phys. Chem.* **31**, 579–601.

Jaynes, E. T. (1982) On the rationale of maximum entropy methods. *Proc. IEEE* **70**, 939–952.

Jaynes, E. T. (1982) *Papers on probability, statistics and statistical physics*, Reidel, Dordrecht.

Kagan, A. M., Linnik, Y. M., and Rao, C. R. (1973) *Characterization problems in mathematical statistics*, Wiley, New York.

Kantz, H. and Schreiber, T. (1997) *Nonlinear time series analysis*, Cambridge University Press.

Karlin, S. and Taylor, H.M. (1975) *A first course in stochastic processes*, Academic Press, Boston, § 9.6.

Kay, J.J. (1984) *Self-organization in living systems*, PhD thesis, Systems Design Engineering, University of Waterloo, Ontario.

Kolmogorov, A. N. (1965) Three approaches to the quantitative definition of information. *Prob. Inf. Transm.* **1**, 1–7.

Krieger, W. (1970) On entropy and generators of measure-preserving transformations. *Trans. Am. Math. Soc.* **149**, 453–464.

Krieger, W. (1972) On unique ergodicity. In *Proc. Sixth Berkeley Symp.*, Vol. 2, University of California Press, Berkeley, 327–346.

Kullback, S. and Leibler, R. (1951) On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Laguës, M. and Lesne, A. (2008) *Invariances d'échelle*, 2nd edition, Belin, Paris. English traduction *Scaling*, Springer, Berlin (2011) in print.

Landauer, R. (1961) Irreversibility and heat generation in the computing process. *IBM J. Res. Deb.* **5**, 183–191.

Lebowitz, J. L. (1993) Boltzmann's Entropy and Time's Arrow. *Physics Today* **46**, 32–38.

Lebowitz, J. L. (1993) Macroscopic laws, microscopic dynamics, time's arrow and Boltzmann's entropy. *Physica A* **194**, 1–27.

Ledrappier, F. and Strelcyn, J. M. (1982) A proof of the estimation from below in Pesin's entropy formula. *Erg. Th. Dyn. Sys.* **2** 203–219.

Lempel, A. and Ziv, J. (1976) On the complexity of finite sequences. *IEEE Trans. Inf. Th.* **22**, 75–81.

Lesne, A. (1998) *Renormalization methods*, Wiley, New York.

Lesne A. (2007) Discrete *vs* continuous controversy in physics. *Math. Struct. Comput. Sci.* **17**, 185–223.

Lesne, A. and Benecke, A. (2008) Feature context-dependency and complexity reduction in probability landscapes for integrative genomics. *Theor. Biol. Med. Mod.* **5**, 21.

Lesne, A., Blanc, J. L., and Pezard, L. (2009) Entropy estimation of very short symbolic sequences. *Phys. Rev. E* **79**, 046208.

Leyton, M. (2001) *A generative theory of shape*, Springer, New York.

Lévy, P. (1965) *Processus stochastiques et mouvement brownien*, Gauthier-Villars, Paris. Reprinted by Éditions J. Gabay, Paris.

Li, M. and Vitanyi, P. (1997) *An Introduction to Kolmogorov complexity and its applications*, Springer, New York.

Mahara, H. and Yamaguchi, T. (2010) Entropy balance in distributed reversible Gray-Scott model. *Physica D* **239**, 729–734.

Martin-Löf, P. (1966) The definition of random sequence. *Inform. Contr.* **9**, 602–619.

McMillan, B. (1953) The basic theorems of information theory. *Ann. Math. Stastist.* **24**, 196–219.

Mugur-Schächter, M. (1980) Le concept de fonctionnelle d'opacité d'une statistique. Étude des relations entre la loi des grands nombres, l'entropie informationnelle et l'entropie statistique. *Annales de l'IHP, section A* **32**, 33–71.

Nicolis, G. and Gaspard, P. (1994) Toward a probabilistic approach to complex systems. *Chaos, Solitons & Fractals* **4**, 41–57.

Nicolis, G. and Prigogine, I. (1977) *Self-organization in nonequilibrium systems*, Wiley, New York.

Parisi, G. (2003) Complexity and intelligence. In A. Vulpiani and R. Livi (editors), *The Kolmogorov Legacy in Physics*, Springer, Berlin, 109–122.

Pesin, Y. (1997) *Dimension theory in dynamical systems. Contemporary views and applications*, University of Chicago Press, Chicago.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006) Maximum entropy modeling of sepecis geographic distribution. *Ecological Modelling* **190**, 231–259.

Phillips, S. J. and Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–175.

Prigogine, I. (1967) *Thermodynamics of irreversible processes*, Interscience Publishers, New York.

Rached, Z., Alajaji, F., and Campbell, L. (2001) Rényis divergence and entropy rates for finite alphabet Markov sources. *IEEE Trans. Inf. Theor.* **47**, 1553–1562.

Robert, C. (1990) An entropy concentration theorem: applications in artificial intelligence and descriptive statistics. *J. Appl. Prob.* **27**, 303–313.

Ruelle, D. P. (1978) *Thermodynamic formalism*, Addison-Wesley, New York.

Ruelle, D. P. (2003) Extending the definition of entropy to nonequilibrium steady states. *Proc. Natl. Acad. Sci. USA* **100**, 3054–3058.

Samengo I. (2002) Estimating probabilities from experimental frequencies. *Phys. Rev. E* **65**, 046124.

Sanov, I. N. (1957) On the probability of large deviations of random variables (in Russian), *Mat. Sbornik* **42**,11–44. (English translation in *Selected Translations in Mathematical Statistics and Probability I* 1961, 213–244, Institute of Mathematical Statstics, Providence.)

Sagawa, T. and Ueda, M. (2009) Minimal energy cost for thermodynamic information processing: measurement and information erasure. *Phys. Rev. Lett.* **102**, 250602.

Schrödinger, E. (1944) *What is life? The physical aspect of the living cell*, Cambridge University Press.

Schulman, L. S. (2010) We know why coffee cools. *Physica E* **42**, 269–272.

Shannon, C. (1948) A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423.

Shinner, J. S., Davison, M., and Landsberg, J. T. (1999) Simple measure for complexity. *Phys. Rev. E* **59**, 1459–1464.

Sinai, Ya. G. (1959) On the concept of entropy for dynamical systems. *Dokl. Acad. Nauk SSSR***124**, 768–771 (in Russian).

Sokal, A. D. (1997) Monte Carlo methods in statistical mechanics: Foundations and new algorithms. In C. C. DeWitt-Morette and A. Folacci (editors), *Functional Iintegration: basics and applications* (1996 Cargèse summer school), Plenum Press, New York.

Sokal, A. D. and Thomas, L. E. (1989). Exponential convergence to equilibrium for a class of random-walk models. *J. Stat. Phys.* **54**, 797–828.

Solomonoff, R. (1978). Complexity-based induction systems: comparisons and con- vergence theorems. *IEEE Trans. Inf. Th.* **24**, 422–432.

Szilard, L. (1929) Uber die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. (On the lessening of entropy in a thermodynamic system by interference of an intelligent being, in German). *Z. Physik* **53**, 840–856.

Touchette, H. (2009) The large deviation approach to statistical mechanics. *Phys. Rep.* **478**, 1–69.

Tribus, M. and McIrvine, E. C. (1971)Energy and information.*Sci. Am.* **225**, 179–188.

Van Campenhout, J. M. and Cover, T. M. (1981) Maximum entropy and conditional entropy. *IEEE Trans. Inf. Th.* **27**, 483–489.

Vovk, V. and Shafer, G. (2003) Kolmogorovs contributions to the foundations of probability. *Prob. Inf. Transm.* **39**, 21–31.

Werhl, A. (1978) General properties of entropy. *Rev. Mod. Phys* **50**, 221–261.

White, H. (1993) Algorithmic complexity of points in dynamical systems. *Erg. Th. Dyn. Sys.* **13**, 807–830.

Wyner, A. D. and Ziv, J. (1989) Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inf. Th.* **35**, 1250–1258.

Ziv, J. and Lempel, A. (1977) A universal algorithm for sequential data compression. *IEEE Trans. Inf. Th.* **23**, 337–343.

Ziv, J. and Lempel, A. (1978) Compression of individual sequences by variable rate coding. *IEEE Trans. Inf. Th.* **24**, 530–536.

Zuk, O., Kanter, I., and Domany, E. (2005) Aymptotics of the entropy rate for a hidden Markov process. *Proc. DCC'05* 173–182. The entropy of a binary hidden Markov process. *J. Stat. Phys.* **121**, 343–360.

Zurek, W. H. (1984) Maxwell's Demon, Szilard's engine and quantum measurements. In G. T. Moore and M. O. Scully (editors), *Frontiers of nonequilibrium statistical physics*, Plenum Press, New York, 151–161.