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

ANNICK LESNE

Mathematical Structures in Computer Science / Volume 24 / Special Issue 03 / June 2014 / e240311
DOI: 10.1017/S0960129512000783, Published online: 28 March 2014

Math. Struct. in Comp. Science (2014), vol. 24, iss. 3, e240311, 63 pages. © Cambridge University Press 2014
doi:10.1017/S0960129512000783

Shannon entropy: a rigorous 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
and
Institut des Hautes Études Scientifiques
35 route de Chartres, F-91440, Bures-sur-Yvette, France
Email: lesne@ihes.fr

Received 18 January 2011; revised 20 October 2011

Statistical entropy was introduced by Shannon as a basic concept in information theory measuring the average missing information in a random source. Extended into an entropy rate, it gives bounds in coding and compression theorems. In this paper, I describe how statistical entropy and entropy rate relate to other notions of entropy that are relevant 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) and statistical physics (Boltzmann entropy). Their mathematical foundations and correlates (the entropy concentration, Sanov, Shannon–McMillan–Breiman, Lempel–Ziv and Pesin theorems) clarify their interpretation and offer a rigorous basis for maximum entropy principles. Although often ignored, these mathematical perspectives give a central position to entropy and relative entropy in statistical laws describing generic collective behaviours, and provide insights into the notions of randomness, typicality and disorder. The relevance of entropy beyond the realm of physics, in particular for living systems and ecosystems, is yet to be demonstrated.

1. Introduction

Historically, many notions of entropy have been proposed. The first use of the word *entropy* dates back to Clausius (Clausius 1865), who coined this term from the Greek *tropos*, meaning transformation, and the prefix *en*- to recall its inseparable (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).

[†] Part of this paper was presented during the 'Séminaire Philosophie et Mathématiques de l'École normale supérieure' in 2010 and the 'Séminaire MaMuPhy' at Ircam in 2011.
This work has been funded by the Agence Nationale de la Recherche, SYSCOMM program, grant DISCO 09-SYSC-003 and by the Institut National de la Santé et de la Recherche Médicale, grant MICROMEGAS PC201104.

It is formally similar to the Boltzmann entropy associated with the statistical description of the microscopic configurations of many-body systems and the way it accounts for their macroscopic behaviour (Honerkamp 1998, Section 1.2.4.; Castiglione *et al.* 2008). The work of establishing the relationships between statistical entropy, statistical mechanics and thermodynamic entropy was begun by Jaynes (Jaynes 1957a; Jaynes 1957b; Jaynes 1982b).

Starting from what was initially a totally different perspective, a notion of entropy rate was developed 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 with so many meanings just misleading?

Thirty years ago, Jaynes gave a short historical account of the different notions of entropy in Jaynes (1980). In the current paper, I propose to give a more detailed overview of the relationships between the different notions of entropy as they appear today, rather than from a historical perspective, and to highlight the connections between probability, information theory, dynamical systems theory and statistical physics. I will base my presentation on mathematical results related to Shannon entropy, relative entropy and entropy rate, which offer a reliable guide, both qualitatively and quantitatively, for the proper use and interpretation of these concepts. In particular, they provide a rationale, as well as several caveats, for using what is called 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* (Shannon 1948) is defined by

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

Shannon entropy quantifies the *unevenness* in the probability distribution p. In particular, the minimum $H(X) = 0$ is reached for a constant random variable, that is, a variable with a determined outcome, which is reflected in a fully localised probability distribution $p(x_0) = 1$ and $p(x) = 0$ for $x \neq x_0$. At the opposite extreme, $H(X)$ is maximal, equal to $\log_2(|\mathcal{X}|)$, for a uniform distribution. An alternative notation for the Shannon entropy is $S(p)$, which underlines the fact that entropy is a feature of the probability distribution p. Entropy does not depend on the graph $x \to p(x)$, that is, it is not a feature of the random variable itself, but only of the set of its probability values. This property is reflected in a permutation invariance of $H(X)$: if we let the variable $\sigma.X$ be obtained by a permutation of the states, that is, labelling states $x \in \mathcal{X}$ by an index i,

$$\mathrm{Prob}(\sigma.X = x_{\sigma(i)}) = p(x_i),$$

then $H(X) = H(\sigma.X)$. Entropy trivially increases with the number of possible states: for an unbiased coin,

$$H = \log_2 2 = 1$$

but for an unbiased dice,

$$H = \log_2 6 > 1.$$

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

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

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

(iii) if we group

$$\begin{aligned} y_1 &= \{x_1, \ldots, x_{k_1}\} \\ y_2 &= \{x_{k_1+1}, \ldots, x_{k_1+k_2}\} \\ &\vdots \\ y_m &= \{x_{n-k_m+1}, \ldots, x_n\} \end{aligned}$$

so that

$$q_i = \sum_{i=k_1+\ldots+k_{(i-1)}}^{i=k_1+\ldots+k_i-1} p_l$$

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

$$\mathcal{H}(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log_2 p_i,$$

[†] 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 don't 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'".

which is *unique* up to the multiplicative constant K. Statistical entropy, which we will denote by $H(X)$ or $S(p)$ from now on, is thus almost uniquely defined by the above natural prescriptions (i)–(iii). It is easy to check from definition (1) that the entropy of a compound of independent events $Z = X_1 \dots X_n$ such that $p_Z = p_{X_1} \dots p_{X_n}$ is simply

$$H(Z) = \prod_{i=1}^{n} H(X_i).$$

Another important property is *entropy convexity*. If we let

$$p = \lambda p^0 + (1-\lambda)p^1,$$

then

$$S(p) \geqslant \lambda S(p^0) + (1-\lambda)S(p^1)$$

or, equivalently,

$$H(X) \geqslant \lambda H(X^0) + (1-\lambda)H(X^1)$$

where X, X^0 and X^1 are random variables with distributions p, p^0 and p^1, respectively. 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 to quantify the information loss in transmitting a given message in a communication channel (Shannon 1948). A noticeable aspect of Shannon's approach is that he ignored semantics and focused on the physical and statistical constraints limiting the transmission of a message, irrespective of its meaning. The source generating the inputs $x \in \mathcal{X}$ is characterised by the probability distribution $p(x)$. Shannon introduced the quantity

$$I_p(x) \equiv -\log_2 p(x)$$

as a measure of the *information* given by the observation of x knowing the probability distribution p. In plain language, it could correctly be said (Balian 2004) that $I_p(x)$ is the surprise in observing x given some prior knowledge of the source summarised 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 1957a; Jaynes 1957b). 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 Section 5.2, which is devoted to data compression and coding theorems). $I_p(x)$ is now normally denoted by $I(x)$, which, regrettably, ignores the essential connection to the distribution p.

The actual message is just one message selected from a set of possible messages, and information is produced when a single 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 classic example is the quantification of the information needed to communicate a play by Shakespeare, depending on whether the receiver knows in advance that he will receive one of the plays by Shakespeare (in which case transmitting only the few first words is sufficient) or not (in which case the whole text of the play has to be transmitted). What changes between the two situations is the *a priori* knowledge, that is, the *set of possible messages*. In the above formalisation, it is described through the *a priori* probability $p(x)$ describing the source. It should thus be emphasised 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 $p(x)$. Information is not an absolute notion, but rather a highly subjective and relative one. 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 is often mixed up with the loose meaning of information in everyday language. From now on, we shall use the term *knowledge* instead of information when the latter is used with its non-technical everyday (plain language) meaning.

Shannon information and its statistical average, Shannon entropy, should not be confused with *Fisher information*, which appears in parametric estimates, that is, the estimate of a parameter a of a probability distribution $p_a(x)$. It is defined as

$$I_F(a) = \int([\partial \ln p_a(x)/\partial a)]^2 p_a(x)dx \tag{2}$$

(for a one-dimensional parameter a; for the multivariate extension, see, for example, Amari and Nagaoka (2000) and Cover and Thomas (2006)). Its main interest comes from the *Cramer–Rao bound* (Kagan *et al.* 1973), which relates Fisher information and the variance $\mathrm{Var}(\hat{a})$ of the estimated value $\hat{a}$ of the parameter a through the inequality

$$\mathrm{Var}(\hat{a}).I_F(a) \geqslant 1.$$

We shall say more about the geometric meaning of Fisher information in relation to relative entropy in Section 2.4.

2.3. *Conditional entropy, relative entropy and the Kullback–Leibler divergence*

Shannon entropy can be extended (Gray 1990; Cover and Thomas 2006) to multivariate random variables. It involves their joint distribution: for example, for two random variables X and Y, it involves taking their values in two *a priori* different (discrete and finite) spaces $\mathcal{X}$ and $\mathcal{Y}$, so we have

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

From this joint entropy, we then define the *conditional entropy*

$$H(X \mid Y) \equiv H(X,Y) - H(Y),$$

which appears to be the average (over Y) of the entropies of the conditional probability distribution $p(X \mid Y = y)$:

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

Using a convexity argument (Honerkamp 1998, Section 1.2.4.), it is straightforward to show that

$$H(X \mid Y) \leqslant H(X) \leqslant H(X,Y).$$

In particular, $H(X|Y) \leqslant H(X)$ reflects the fact that the uncertainty about X is never increased by the knowledge of Y. In the case of multiple conditioning, we have

$$H(X|Y,Z) \leqslant H(X|Y) \leqslant H(X).$$

When the random variables X and Y have the same state space $\mathcal{X}$ and distributions p_X and p_Y, respectively, we can 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 $S_{rel}(p_X|p_Y) \leqslant 0$. 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 $\mathcal{X}$,

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

$D(p||q)$ is not a distance since it is not symmetric and does not satisfy the triangle inequality; the only property it shares with a distance is that $D(p||q) \geqslant 0$, with $D(p||q) = 0$ if and only if $p = q$. Nevertheless, we shall see that it has a useful geometric interpretation and some useful properties (Section 2.4 and Section 3.2).

To give an illustration of the use and interpretation of these quantities, let us consider a stationary Markov chain $(X_t)_{t \geqslant 0}$. 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 (where we use p_t to denote the distribution at time t and p_{stat} to denote the stationary distribution). It is important not to confuse:

— the conditional entropy

$$H(X|Y) = H(X,Y) - H(Y)$$

of the random variables X and Y, which could take their values in different sets $\mathcal{X}$ and $\mathcal{Y}$. Its computation requires us 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, the Kullback–Leibler divergence $D(p_X||p_Y)$ between their probability distributions, which are both defined on $\mathcal{X}$.

The distinction between relative and conditional entropies becomes even clearer when we introduce the *mutual information*:

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

of two random variables X and Y. This 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 and p_Y are 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, and it equals $H(X)$ if $X = Y$. This notion can be extended into a *conditional mutual information* (mutual information between X and Y given Z) defined by

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

2.4. *Information geometry*

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

$$dp^2(a) = \sum_{i,j} g_{i,j}(a) da^i da^j \tag{10}$$

where

$$g_{ij}(a) = \int \frac{\partial \log_2 p(x,a)}{\partial a_i} \frac{\partial \log_2 p(x,a)}{\partial a_j} p(x,a)dx,$$

so

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

This metric endows the parametrised 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 (for example, the state space) are discrete. This makes the tools of differential geometry available for the study of statistical structures as geometrical structures.

The components of the parameter a now have to be estimated from the data. As mentioned above for $d = 1$, a bound on the estimator is given by the *Cramer–Rao theorem*, which states (Kagan *et al.* 1973) that $V_a(\hat{a}) - G(a)^{-1}$ 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 $p(.,a)$. This 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 making estimates from experimental data and for control.

2.5. *Behaviour of entropy for coarse graining and local averaging*

The compositional law requirement (iii) involved in the construction of Shannon entropy (see Section 2.1) explicitly expresses the behaviour of Shannon entropy when there is *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 $Y = y$, that is, the conditional entropy of X knowing $Y = y$ averaged over the probability distribution of Y. This additional term is just the conditional entropy $H(X \mid Y)$), from which it follows that

$$H(X) = H_{cg}(Y) + H(X \mid Y). \tag{12}$$

Now $H_{cg}(Y) \leqslant H(X)$, with strict inequality when the coarse graining is non-trivial, that is, when $m < n$.

On the other hand, Shannon noted in his seminal 1984 paper that any change towards equalisation of the probabilities $p_1, \dots, p_n$ increases H. In particular, such a change is achieved through a *local averaging*. For instance, defining

$$p'_i = [p_i + p_{i+1}]/2$$

where

$$p'_n = (p_n + p_1)/2,$$

we get

$$H_{av} = S(p') \geqslant H = S(p).$$

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, and this leads to an entropy decrease

$$H_{cg} \leqslant H,$$

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. From 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 then be said to decrease entropy. A second way to model the situation is to describe a noisy channel by 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 $H(X) + H(Y)$, which is itself larger than the entropy $H(X, Y)$ of the source, so in this case, the truncated transmission corresponds to an increase in entropy. Entropy is thus extremely sensitive to the set of possible events considered, here $\mathcal{X}$ or $\mathcal{X} \times \mathcal{Y}$, respectively. This example underlines the irrelevance of talking about information loss or information transfer between systems having different states spaces. Here again, we recall the caveat that speaking of (missing) information makes sense only with respect to our prior knowledge of the possible states.

2.6. *Continuous extension*

The extension of entropy to continuous-valued random variables was discussed in Shannon (1948) and is now a textbook matter (Ihara 1993). In this case, the entropy expression is

$$S(p) = -\int_{\mathcal{X}} p(x) \log_2 p(x) dx$$

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 under a change of coordinate $y = f(x)$, leading us to replace $p(x)dx$ by $q(y)dy$ with $p(x) = |f'(x)|q(f(x)))$. While 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 under a change of coordinates[†]. Here we see an instance of the general fact that continuous probabilities fundamentally require more delicate handling, and can lead to well-known paradoxes, such as 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 to bear in mind is that the meaningful quantity having a proper mathematical behaviour (for example, under a change of coordinates) is not $p(x)$, but $p(x)dx$.

[†] In fact, $D(\mu||\mu_0)$ is defined for any pair of probability measures on a Polish topological space $\mathcal{X}$ (for example, a closed subset of $\mathbf{R}^d$), provided the probability measure μ is absolutely continuous with respect to μ_0. In that case, we have $D(\mu||\mu_0) = \int_{\mathcal{X}} d\mu \log(d\mu/d\mu_0)$; otherwise $D(\mu||\mu_0) = +\infty$.

3. Concentration theorems and maximum entropy principles

3.1. *Types and entropy concentration theorems*

In this section we consider sequences $\bar{x}_N \in \mathcal{X}^N$ of N independent and identically distributed random variables with values in a finite set $\mathcal{X}$. In current terminology, this space $\mathcal{X}$ is called the *alphabet* and its elements *symbols*, in reference to messages in communication theory. The definitions and results of this section will apply equally 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 $\bar{x}_N$, which 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 $\bar{x}_N$, and is thus an observable quantity that can be derived from an observed sequence $\bar{x}_N$ as the *normalised 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 called ‘types’. Each probability distribution p on $\mathcal{X}$ defines a type, that is, a subset of $\mathcal{X}^N$. The space $\mathcal{X}^N$ can be thought of as a microscopic phase space (see Section 8.2); the types then define macrostates. There are at most $(1+N)^{|\mathcal{X}|}$ different types (Cover and Thomas 2006), while the number of sequences in $\mathcal{X}^N$ grows exponentially. In consequence, at least one type has exponentially many elements (Csiszár 1998). In fact, we shall see that, asymptotically, one type contains most of the elements. In its simplest formulation, the *entropy concentration theorem* states (Georgii 2003) that

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

(where Card denotes the cardinal), which can be extended to the relaxed statement, for any sequence p_N tending to p as $N \to \infty$,

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

Using Prob to denote the equiprobable distribution on the microscopic phase space $\mathcal{X}^N$ (that is, the normalised cardinal), this statement can be rewritten to give

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

This can accommodate some fixed tolerance ϵ, namely, using $|q-p|$ to denote any distance (such as the quadratic distance) between the probability distributions q and p on $\mathcal{X}$,

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

Let p^* be the distribution maximising Shannon entropy, so

$$H(p^*) = \log_2 |\mathcal{X}|$$

while

$$H(p) - \log_2 |\mathcal{X}| < 0$$

for any other type $p \neq p^*$. This means that, asymptotically, the type of p^* contains almost all sequences. More precisely, it can be shown that

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

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

$$\begin{aligned} \lim_{N\to\infty} \text{Prob}[\bar{x}_N \in \mathcal{X}^N, p_{\bar{x}_N} = p] = 0 \\ \lim_{N\to\infty} \text{Prob}[\bar{x}_N \in \mathcal{X}^N, |p_{\bar{x}_N} - p^*| > \epsilon] = 0. \end{aligned} \tag{18}$$

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

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

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

The statistical average $\langle a_\alpha(X)\rangle$ computed with respect to the empirical distribution $p_{\bar{x}_N}$ is just 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) satisfies the constraints, that is, whether its type belongs to $\mathcal{D}$. Using $p^*_{\mathcal{D}}$ to denote the distribution maximising 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:

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

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

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

and replacing $p^*_{\mathcal{D}}$ with the distribution $p^*_{\mathcal{D}_\delta}$ maximising Shannon entropy in $\mathcal{D}_\delta$. It can be shown (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, the question of a concentration theorem for correlated sequences arises: we shall describe such an extension in Section 5.1, where we show that 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 might also want to make a statement about the *asymptotic weight* of the different types, given the distribution p_0 of the elements. Accordingly, we shall now consider sequences of independent elements whose states are identically distributed according to the distribution p_0 on $\mathcal{X}$ (sometimes called the *reference distribution*). Since the sequences in $\mathcal{X}^N$ are uncorrelated, their probability distribution is just 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 behaviour is no longer the entropy, but the relative entropy, or, equivalently, its opposite, the Kullback–Leibler divergence (6):

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

This gives us (15) again when p_0 is uniform; indeed, using p_{unif} to denote the equiprobable distribution on $\mathcal{X}$, we get

$$D(p||p_{unif}) = \log_2 |\mathcal{X}| - H(p).$$

We can also accommodate some fixed tolerance $\epsilon > 0$:

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

This contains 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 us to control the remainder: that is, large deviations, finite-size errors in the estimate 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 modifying the reference distribution p_0 in the least biased way. Considering the same constrained sets $\mathcal{D}_\delta$ and $\mathcal{D}$ as in Section 3.1, the solution is given by the closest distributions to p_0, as measured by the Kullback–Leibler divergence, that belong to $\mathcal{D}_\delta$ and $\mathcal{D}$, respectively. 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, it can be proved (Robert 1990) that:

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

(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 $\mathcal{D}_\delta$: that is, for any neighbourhood $\mathcal{V}_\delta$ of $p^*_{\mathcal{D}_\delta}$ in $\mathcal{D}_\delta$ (for the narrow topology, that is, the weak topology for bounded continuous real functions on $\mathcal{X}$), we have $\exists\alpha > 0$, $\exists N_0$ such that $\forall N \geqslant N_0$, we have

$$p_0^{\otimes N}[\bar{x} \in \mathcal{X}^N, p_{\bar{x}_N} \notin \mathcal{V}_\delta \mid p_{\bar{x}_N} \in \mathcal{D}_\delta] \leqslant e^{-N\alpha}.$$

(iv) $p^*_{\mathcal{D}}$ has the concentration property in $\mathcal{D}$: that is, for any neighbourhood $\mathcal{V}$ of $p^*_{\mathcal{D}}$ in $\mathcal{D}$, we have $\exists\alpha > 0$, $\exists N_0$ such that $\forall N \geqslant N_0$, we have

$$p_0^{\otimes N}[\bar{x} \in \mathcal{X}^N, p_{\bar{x}_N} \notin \mathcal{V} \mid p_{\bar{x}_N} \in \mathcal{D}] \leqslant e^{-N\alpha}.$$

This is a *large deviation* result (Ellis 1985; Touchette 2009), stated in a way that supplements the previous concentration theorem in Section 3.1 since it allows us to control the remainder. It is valid only with the Kullback–Leibler divergence: other measures of the distance between the two distributions, for example, the quadratic distance, do not satisfy a large deviation statement (Robert 1990). Note that

$$D(p^*_{\mathcal{D}_\delta}||p_0) \leqslant D(p^*_{\mathcal{D}}||p_0)$$

since $\mathcal{D} \subset \mathcal{D}_\delta$. We shall now consider the distribution σ^* maximising the relative entropy $-D(\sigma||p_0)$ in the complement of $\mathcal{V}$ in $\mathcal{D}$, so, by construction,

$$D(\sigma^*||p_0) > D(p^*_{\mathcal{D}}||p_0).$$

The exponent α is given roughly by

$$D(\sigma^*||p_0) - D(p^*_{\mathcal{D}}||p_0).$$

The larger $\mathcal{V}$ is, the more distant σ^* is from $p^*_{\mathcal{D}}$, the larger $D(\sigma^*||p_0)$ is, and hence the larger is its difference form $D(p^*_{\mathcal{D}}||p_0)$, and the larger α is. This 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 faster when $\mathcal{V}$ is larger, that is, 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 normalised cardinal and we recover the specific 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):

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

This is a large deviation result (Ellis 1985; Touchette 2009), which could also be thought of as a projection (Georgii 2003) – see Section 3.3. The Sanov theorem can be extended to continuous densities, where the space $\mathcal{X}$ becomes 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 some constraint

$$\phi(x) \in \mathcal{C} \Longleftrightarrow \int_{\mathcal{X}} a(x)\phi(x)dx = A. \tag{27}$$

where we have used $a(x)$ to denote some field and A to denote some fixed number. 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 maximises the relative entropy under the constraint of belonging to $\mathcal{C}$.

3.3. *Geometric interpretation*

The Kullback–Leibler divergence is a useful tool, and is suitable for use 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

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

The distributions in $\mathcal{C}$ minimising $D(.||p_0)$ could be called the 'orthogonal projections' of p_0 onto $\mathcal{C}$. When $\mathcal{C}$ is closed for the weak topology, such minimisers are guaranteed to exist (Georgii 2003). If, moreover, $\mathcal{C}$ is convex, the minimiser $Q(p_0)$ is uniquely determined (Csiszár 1975), and 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, the Kullback–Leibler divergence could be more natural and more efficient than true functional distances, such as the quadratic distance. 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 $(p_k)_{k=1,...,m}$, the Kullback–Leibler divergence allows us to define a notion of *empirical average* $\bar{p}$ of the sample by

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

(Georgii 2003; Balding *et al.* 2008). This definition of an average object is suitable when the 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 the Kullback–Leibler divergence is related to *conditional expectation.* Let Σ be a sub-sigma algebra of the original Σ_0. Let p be the original probability density of the random variable X being considered and 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, that is, a coarse description of the random variable X. An explicit computation using

$$\int p(x)\log_2[E^{\Sigma}p(x)]dx = \int E^{\Sigma}p(x)\log_2[E^{\Sigma}p(x)]dx$$

according to the above definition, leads straightforwardly to the relation

$$D(p||E^{\Sigma}p) = S(E^{\Sigma}p) - S(p) \geqslant 0. \qquad (31)$$

Moreover, we get

$$D(p||q) - D(p||E^{\Sigma}p) = D(E^{\Sigma}p||q)$$

for any Σ-measurable density q, and thus (Csiszár 1975)

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

3.4. *Maximum-entropy inference of a distribution*

An accepted heuristic principle in constructing a statistical model given some prior knowledge (for example, experimental data) is to minimise 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 of the observed process, we should consider equiprobable outcomes. This principle dates back to the *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 (for example, a condition on its support and/or prescribed values for some of its moments), a constructive method for implementing this principle is to maximise the Shannon entropy under the constraints. Reconstruction of the probability distribution using the maximum entropy principle is by no means restricted to statistical mechanics, or any other specific applications, but is a general method of inference under *a priori* constraints, ensuring that no additional arbitrary assumptions, that is, no biases, are introduced (Frank 2009). Any discrepancy between predictions and observations would presumably be due to an ill-constrained maximum entropy principle, and provides evidence supporting a need for additional constraints (or the need to relax spurious constraints). Constraints amount to restricting 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 because of 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

$$\langle a_\alpha(X)\rangle_p = \sum_{i=1}^{n} p(x_i)a_\alpha(x_i), \qquad \alpha = 1,\ldots,m,$$

and we want to estimate another average $\langle b(X)\rangle$ (which is precisely the issue encountered in statistical mechanics – see Section 8). As just explained, the problem can be reformulated as follows. What is the distribution $p(x)$ on the finite space $\mathcal{X} = \{x_1,\ldots,x_n\}$ maximising Shannon entropy under the following constraints:

(i) $p(x) \geqslant 0$ for any $x \in \mathcal{X}$;
(ii) $\sum_{x\in\mathcal{X}} p(x) = 1$;
(iii) for $\alpha = 1,\ldots,m$,

$$\sum_{x\in\mathcal{X}} p(x)a_\alpha(x) = A_\alpha.$$

The solution (Jaynes 1982a; Frank 2009) is

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

where the Lagrange multipliers λ_α are determined by the need to satisfy the constraints (iii) and the multiplicative constant C ensures the proper normalisation given by (ii). This solution had already been established by Shannon for continuous distributions in some specific contexts (Shannon 1948): the distribution on $[-\infty,\infty[$ maximising entropy at fixed average μ and fixed variance σ^2 is the Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$; and the distribution on $[0,\infty[$ maximising entropy at fixed average μ is the exponential distribution

$$f(x) = (1/\mu)e^{-x/\mu}.$$

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

Prior knowledge about the system in the absence of constraints is expressed in 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 maximising the relative entropy $S_{rel}(p|p_0)$ under constraints, or, equivalently, minimising 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 the maximum entropy principle applies to discrete distributions, the maximum relative entropy principle can also be applied to continuous distributions without any ambiguity (see Section 2.6). The density maximising relative entropy under the constraint $\langle a(X)\rangle = A$ is given (up to a normalisation factor) by (Kagan *et al.* 1973)

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

and λ is such that

$$\int_{\mathcal{X}} a(x)f_\lambda(x)dx = A. \tag{35}$$

A similar statement holds with an array of conditions, in which case $a(x)$ and λ are vectors, and the product has to be replaced by a scalar product $\lambda^t.a(x)$. A justification of the maximum relative entropy principle was given in Van Campenhout and Cover (1981): the distribution maximising the relative entropy can be characterised as the limit of a sequence of conditional probabilities. We shall now consider independent and identically distributed variables $X_1, \ldots, X_N$ with density p_0. The conditional probability of X_1 under the constraint

$$(1/N)\sum_{i=1}^{N} a(X_i) = A$$

on the empirical average converges to the distribution maximising the relative entropy, namely,

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

The usefulness of this theorem depends largely on the proper identification of the observable $a(x)$. The ideal situation is to identify a function $h(x)$ such that

$$\bar{a}_N = (1/N)\sum_{i=1}^{N} a(X_i)$$

is a sufficient statistics summarising all the relevant knowledge about the sample at the macroscopic scale. In this case, given $\bar{a}_N$, 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 principle. Specifically, starting from a reference distribution p_0, we could first determine the maximum relative entropy distribution p_1^* associated with a set of constraints

$$\langle a_i\rangle = A_i, \qquad i = 1, \ldots, n.$$

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

$$\langle a_i\rangle = A_i, \qquad i = n+1, \ldots, n+m.$$

This would then be the same as determining the maximum relative entropy distribution associated with the set of constraints

$$\langle a_i\rangle = A_i, \qquad i = 1, \ldots, n+m$$

starting from the reference distribution p_0.

A serious conceptual problem in the practical application of the maximum entropy inference method was pointed out in Haegeman and Etienne (2010): the distribution obtained by maximising entropy seems to depend on the chosen setting. 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. From one 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 N^M possible configurations $\mathbf{m} = (m_1, \ldots, m_N)$. From a second viewpoint, the M cells are labelled but the individuals are now indistinguishable, 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 through

$$n_m = \sum_{i=1}^{N} \delta(m, m_i).$$

The maximum entropy principle applied to an 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, the maximum entropy principle applied to an 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 taking into account the identity of the individuals is a piece of 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 is always smaller than the entropy computed at a more refined level). This leads to the open question of the proper *a priori* choices to be made in using the maximum entropy inference method, since the choice of the configuration space has a strong influence on the resulting distribution and the macroscopic quantities (averages and moments) that can be computed from it. The paradoxes are solved by a consistency argument (constraints must be consistent with the distribution under consideration) or by a return to the mathematical foundations (types and the entropy concentration theorems). Here, the concentration theorems only apply 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 the maximum entropy principle to the reconstruction of either the distribution $p(\mathbf{m})$ or $p(\mathbf{n})$.

It should be stressed that the maximum entropy principle is justified not only as the formalisation of the intuitive indifference principle of Laplace, but also, rigorously, by the entropy concentration theorems of Section 3.1. These theorems state that, asymptotically (that is, 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. The experimental observation of a microscopic configuration yielding a different empirical distribution $p_{obs} \neq p^*$ has an exponentially small probability, which decreases like $2^{-[H(p^*)-H(p_{obs})]}$. This statement can be extended to constrained distributions: the configurations whose type satisfies a set of linear constraints concentrate about the distribution maximising relative entropy under the same constraints. Accordingly, almost all microscopic configurations behave in the same way with respect to 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, thus replacing Shannon entropy $H(p)$ with the relative entropy $-D(p||p_0)$. The asymptotic nature of the statement, which is reflected in a condition on the sample size N, could nevertheless be limiting in some situations. Moreover, it relies on the independence of the different elements or individuals, which often cannot be assumed. We shall see an extension of the concentration theorems to correlated populations or samples in Section 5.1, which involves an average entropy rate h instead of the entropy or relative entropy. We stress that maximum entropy prediction only applies to the empirical distribution or type, that is, 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 know 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 maximising the entropy of the probability distribution, namely, $(p_k)_{k=1,\ldots,6}$, under the constraint

$$\sum_{k=1}^{6} k p_k = \alpha$$

on its average and the normalisation constraint

$$\sum_{k=1}^{6} p_k = 1.$$

This gives

$$p_k = \exp(\lambda_0 + k\lambda_1)$$

where λ_0 and λ_1 are chosen to satisfy the normalisation and average constraints. The rationale for using the maximum entropy principle here comes from the theory of types and the concentration theorems (see Section 3.1): for any fixed ϵ, the asymptotic behaviour for the conditional probability is given by

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

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 maximising the entropy at fixed average α. Prob corresponds to equiprobable configurations: specifically, in the case of dice, it is given by the cardinal divided by the total number 6^N of configurations (a uniform reference distribution).

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

A corollary of the Shannon maximum entropy principle is the *Burg maximum entropy theorem*, which states that the process that maximises entropy subject to correlation constraints is an appropriate autoregressive Gaussian process (Cover and Thomas 2006). More explicitly, the stochastic process maximising 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 centred 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 least 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. Dedicated hypothesis testing should be developed here to check whether the underlying process is indeed linear, or not.

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

In this section we discuss a possible extension of the method of types to uncorrelated random graphs – this belongs to the expanding field of the 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 (that is, $A_{ij} = 1$ if there is an edge between the nodes i and j, otherwise $A_{ij} = 0$, 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 then deduce the normalised histogram $p_A(k)$ of this sequence, which is just the empirical degree distribution. The average degree $\langle k \rangle_{p_A}$ with respect to this empirical distribution p_A coincides with the degree average $\sum_{i=1}^{N} k_i/N$, which is itself equal to $2M(A)/N$ where $M(A)$ is the number of edges of the graph. These 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 full knowledge about the graph, at the level of the pairs 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 under 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 a uniform probability distribution

  $$\mathcal{Q}^N_{micro}(M_0) = 1/|\mathcal{E}^N_{micro}(M_0)|$$

  (we can use some tolerance δM to 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 and independently according to the distribution p_0. The degree sequence $[k]$ is thus a realisation 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 thought of 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 use $\mathcal{N}_N(p)$ to denote the number of sequences of length N having the type $p \in \mathcal{P}(\mathbf{N})$. The Csiszár–Körner theorem (Csiszár and Körner 1981) then states that for any sequence $(p_N)_N$ such that $\lim_{N\to\infty} p_N = p_0$, we have

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

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 $(X_t)_{t\geqslant 0}$ (in discrete time t) with values in a finite set $\mathcal{X}$, the Shannon entropy of the array $(X_1, \ldots, X_n)$ is called the *block entropy* of order n and denoted by H_n. This 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 quantitatively captures all the correlations having a 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 (which we shall call 'symbols' from now on). Note that n-words and the associated block-entropy should not be confused with coarse graining or local averaging – see Section 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, the definition and properties of the conditional entropy (Karlin and Taylor 1975, Section 9.6) give us

$$\begin{aligned} 0 &\leqslant H(X_{n+1} \mid X_1, \ldots, X_n) \\ &\leqslant H(X_{n+1} \mid X_2, \ldots, X_n) \\ &= H(X_n \mid X_1, \ldots, X_{n-1}). \end{aligned} \tag{41}$$

This inequality could be rewritten as

$$\begin{aligned} 0 &\leqslant H_{n+1} - H_n \\ &\leqslant H_n - H_{n-1}, \end{aligned}$$

which implies the existence of the *Shannon entropy rate* (Karlin and Taylor 1975, Section 9.6; Cover and Thomas 2006):

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

where

$$\overleftarrow{X} = (X_i)_{-\infty < i \leqslant -1}.$$

This entropy rate h can be equivalently defined (Karlin and Taylor 1975, Section 9.6; Cover and Thomas 2006) as the limit

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

This limit exists if $\lim_{n\to\infty} H_{n+1} - H_n$ exists, and it then takes the same value; we shall here consider situations where the two limits exist, and thus coincide. h is an asymptotic quantity characterising the global statistical features of the source. In particular, it captures correlations of any range, and thus provides a quantitative measure of the overall temporal organisation of the process. We will use the denotations:

$$\begin{aligned} h_n &= H_{n+1} - H_n = H(X_{n+1}|X_1, \dots, X_n) \\ h_{n,av} &= \frac{H_n}{n}. \end{aligned} \tag{44}$$

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

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

An important point in using the entropy rate for data analysis (Lesne *et al.* 2009) is that h makes sense for both deterministic and stochastic sources. If we consider a sequence $(X_1, \dots, X_n)$ of length n, it can be shown (Karlin and Taylor 1975, Section 9.6) that a random shuffle σ increases entropy, that is,

$$H_n(\sigma.X) \geqslant H_n(X)$$

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 for assessing 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 the interpretation to be used in practical contexts is a whole domain of research, deserving of its own a critical review (Kantz and Schreiber 1997; Lesne *et al.* 2009), and is far beyond the scope of the current paper.

4.2. *Examples and special cases*

For a sequence of independent and identically distributed random variables, $h = H_1$, that is, h reaches its upper bound (at given symbol frequencies). Temporal correlations always reduce h. For a stationary Markov chain of order 1, we have

$$h = H_2 - H_1,$$

while for a stationary Markov chain of order q, we have

$$H_n = H_q + (n - q)h$$

when $n \geqslant q$. In this case, $h_n = h$ exactly when $n \geqslant q$, while $h_{n,av}$ gives only an approximation for 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 $\mathrm{Prob}(B = 1) = p$, and 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 any insights into the underlying process and its features from the value of h itself: only a differential study makes sense, and this could be based on a comparison between two experimental systems (for classification purposes) or between a real system and a model (for quality assessment), or between two models (for model selection purposes). Only $h = H_1$ is directly meaningful, and indicates the absence of temporal correlations in the process. Even $h = 0$ does not lead to a clear-cut conclusion since it may be observed for both a periodic dynamics and a dynamics at the onset of chaos.

We will now consider a more complicated situation, namely, 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 the symmetric transition matrix (46) as above, and E is a Bernoulli noise process with parameter ϵ. The 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) that for small ϵ, we have

$$h(Y) = h(X) + \epsilon c_1 + \mathcal{O}(\epsilon^2)$$

with

$$c_1 = 2(1 - 2p) \log_2[(1 - p)/p].$$

It is notable that $c_1 > 0$ when $0 < p < 1$, so $h(Y) > h(X)$ 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 $H_n(Y) \geqslant h(Y)$, it follows that for δ small enough and n_0 large enough, the inequality

$$H_n(Y) > \delta + H_n(X)$$

holds for any $n \geqslant n_0$. It should also be noted that

$$H_1(X) = H_1(Y) = 1,$$

that is, the noise does not break the symmetry inasmuch 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 the information content and the influence of noise on the transmission cannot be appreciated using the Shannon entropy alone.

4.3. *Information-theoretic interpretation*

The inputs in the information-theoretic interpretation of Shannon entropy H (see Section 2.2) were the elementary states or symbols $x \in \mathcal{X}$. We will now consider the more realistic case where the message is formed by the *concatenation of symbols* emitted successively by the source. For independent symbols, the source is still fully characterised by the entropy H of the elementary distribution. In the general case, time correlations are present between sucessive symbols and we have recourse to h to characterise the source. It is thus important to distinguish H and h: we have

$$h \leqslant H \leqslant \log_2 |\mathcal{X}|$$

and $h < H$ when correlations are present.

Indeed, using the stationarity of the process (Feldman 2002), we have

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

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 it is impossible to disentangle the two contributions. By contrast, H only provides a quantitative characterisation of the unevenness of the probability distribution of the symbols. Using the expression

$$h = \lim_{n\to\infty} H(X_0|X_{-1},\dots,X_{-n}),$$

another interpretation of h is the information required to predict X_0 knowing the whole past.

By definition and the information-theoretic interpretation of Shannon entropy (see Section 2.2), h is the average information given 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 symbol frequency distribution. This means that some redundancy is present in a sequence of length N and, on average,

$$N_{\text{eff}} = Nh/\log_2 |\mathcal{X}|$$

bits are enough to represent the sequence ($N_{\text{eff}} = Nh$ in the case of a binary sequence). The entropy rate also plays a role in statistics in that it captures the time correlations of the process, which is central in controlling the error bars in estimating 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 through

$$2\pi e \sigma_\infty^2 = 2^{2h(X)}.$$

More generally, when making estimates from an experimental sequence,

$$N_{\text{eff}} = Nh(X)/\log_2 |\mathcal{X}|$$

is the effective length of the sequence, which is relevant in appreciating the importance of finite-size effects. The notions of entropy rate $H(X)$ and effective length N_{eff} thus provide the foundations for estimation theorems for correlated samples: for example, in estimating 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 remains a wide range of possible temporal structures for a dynamics characterised by a given entropy rate h. This observation motivated the search for additional measures to derive quantitative characterisations of temporal structures or patterns and their statistical complexity (Feldman and Crutchfield 1998; Feldman 2002). A first direction was 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 behaviour of a statistical measure of complexity, namely one vanishing for regular, for example, periodic, and for fully random distributions (Shinner *et al.* 1999). Nevertheless, this quantity $Q(p)$ contains almost exactly the same knowledge about the distribution p as 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 into $p \to Q(p)$ compared with $p \to H(p)$. Similarly, the quantity $h(h_{max} - h)$ is not a complexity measure since it does not give us any further insights into the structure and organisation of the system compared with the 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 called the *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, for example, for periodic signals and at the onset of chaos. Excess entropy allows us to discriminate the different situations associated with a vanishing entropy by capturing the way H_n/n converges to $h = 0$. For instance, $H_n = \mathit{const}$ for a periodic signal, while $H_n \sim \log_2 n$ at the onset of chaos (Feldman 2002). More generally (Grassberger 1986),

$$H_n \sim E + nh + \text{h.o.}.$$

An equivalent expression for excess entropy is (Badii and Politi 1997)

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

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

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

Using $\theta.X$ to denote the shifted sequence, such that

$$(\theta.X)_t = X_{t+\theta},$$

it can be shown (Blanc *et al.* 2011) that the mutual information rate satisfies

$$\begin{aligned} i(X,\theta.X) &= h(X) \\ i(X,\theta.Y) &= i(X,Y). \end{aligned}$$

Shannon actually introduced the mutual information rate between the input and output signals in the section of his historic 1948 paper devoted to transmission in a noisy channel, and called it the *rate of actual transmission*. Using X to denote 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, that is, the entropy of the message X emitted by the source given the output Y. We have $h(X|Y) = 0$ when knowledge of the output sequence $(y_1,\dots,y_N)$ allows us 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:

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

Another 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 the Rényi divergence of order α

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

which recovers the Shannon entropy and the Kullback–Leibler divergence for $\alpha = 1$ through the equations

$$\lim_{\alpha \to 1} H^{(\alpha)}(p) = H(p)$$

and

$$\lim_{\alpha \to 1} D^{(\alpha)}(p||q) = D(p||q),$$

respectively.

Similarly, the Rényi entropy of order α can be extended 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 these extended entropy rates is to give a tunable weight to rare events: for example, rare events contribute relatively more in $h^{(\alpha)}$ for $\alpha < 1$ than in h. Their drawback is the lack of a subadditivity property except for $\alpha = 0$ and $\alpha = 1$. This is also a problem for non-extensive thermostatistics, which is based on these generalised entropies (Balian 2004; Balian 2005).

4.5. *Spatial extension*

Block-entropies and entropy rates are statistical descriptors of a time series. Technical care is needed to extend these notions and apply them to the quantification of a *spatially extended structure* by considering spatial labels (x, y) or (x, y, z) instead of time t (Grassberger 1986). We can obviously compute the Shannon entropy of the probability distribution describing the fraction of space occupied by the different species forming the pattern (normalised over the different species). A more refined quantification involves Shannon entropies at different observation scales ϵ (local averages). Considering a partition of the space into $N(\epsilon)$ disjoint cells of size ϵ, and using $p_i(\epsilon)$ to denote the measure of the cell i, with

$$\sum_{i=1}^{N(\epsilon)} p_i(\epsilon) = 1,$$

normalised over the different spatial cells, a meaningful index is the *information dimension* (Badii and Politi 1997; Castiglione *et al.* 2008), which describes the ϵ-dependence of the entropy and is given by

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

Another promising but not yet much developed index is the multivariate extension of the entropy rate devised for a time sequence. For this, we have to consider increasing sequences $(B_n)_n$ of multidimensional blocks (for example, squares for a spatial structure in the plane) and

$$h = \lim_{n \to \infty} H_{B_n}/|B_n|.$$

We then check that the resulting entropy rate does not depend on the chosen sequence: that is, if

$$An \subset B_n \subset A_{n+1},$$

then $h([A]) = h([B])$. This would allow us to provide evidence for the existence of patterns, in exactly the same way as the entropy rate h is exploited in time series analysis to provide evidence for non-trivial temporal organisation (Kantz and Schreiber 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.

As far as I am aware, the question as to whether there is any relationship between the degree of spatial order of the structure and pattern and the thermodynamic entropy production when this structure or pattern is associated with a nonequilibrium state of some process (dissipative structure) is still open (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, and thus be associated with the same statistical entropy.

5. Asymptotic theorems and global behaviour of correlated sequences

5.1. *Shannon–McMillan–Breiman theorem*

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

$$\hat{P}_m(\bar{x}) = p_m(x_0, \ldots, x_{m-1}),$$

the asymptotic equipartition property is given by

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

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

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

and

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

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

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

and

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

The asymptotic equipartition property for a sequence of independent and identically distributed variables is simply a 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. The Shannon–McMillan–Breiman theorem extends the law to cover correlated sequences. Nevertheless, all available results apply only to stationary sources, which could be a strong limitation in practical situations.

Another corollary of the Shannon–McMillan–Breiman theorem provides a quantitative description of how h accounts in an effective way for the correlations present within the sequence. Specifically, the effective probability of a new symbol, knowing the sequence of length l that precedes it, is asymptotically (that is, for $l \to \infty$) either e^{-h} or 0 depending on whether the ensuing $(l+1)$-word is typical or not. By contrast, it is equal to the symbol frequency in the case where there are no correlations within the sequence. We thus recover the interpretation of h as 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, Section 9.6).

For N asymptotically large, the Shannon–McMillan–Breiman theorem guarantees that, up to second-order terms, we have $H_N \approx \log_2 \mathcal{N}_N$ where $\mathcal{N}_N$ is the number of (asymptotically equiprobable) typical sequences. We shall see in Section 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 given by the reception of a message (that is, one of these $\mathcal{N}_N$ messages). It is important to note that the Shannon–McMillan–Breiman theorem deals with *probable* sequences, while a *grammar* describes the set of *possible* sequences, or, equivalently, the rules for generating all possible sequences.

Two derived formulations of the Shannon–McMillan–Breiman theorem can be useful. For the first, we 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 formulation was given in Shannon (1948, Theorem 4). For this, we first sort the sequences of length N in order of decreasing probabilities and define $n(q)$ as the number of sequences (starting with the most probable one) needed to accumulate a total probability q (where $0 < q < 1$ is fixed and independent of N). Then

$$\lim_{N\to\infty} (1/N) \log_2 n(q) = h.$$

5.2. *Compression of a random source*

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

In the most general case, the optimal encoding for source compression was introduced by Shannon and is known today as the *Shannon–Fano code*. Given a finite set $\mathcal{X}$ of elements x and their probability distribution $p(x)$, a binary code is a correspondence $x \to w(x)$ where $w(x)$ is a binary word, that is, a finite string of 0s and 1s representing x. We use W to denote the finite set of codewords representing the elements of $\mathcal{X}$, and l_w to denote the length of the codeword w. The code is unambiguously and locally decodable if the following condition, known as *Kraft's inequality*, is satisfied:

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

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 unambiguous and locally decodable character of the code: indeed, if $l'_w \geqslant l_w$ for any word w, then $\Sigma(l') \leqslant \Sigma(l)$. Given the probability $\tilde{p}(w) \equiv p[x(w)]$, minimisation of the average length $\sum_{w\in W} l_w \tilde{p}_w$ at fixed $\Sigma(l) = 1$ gives

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

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

$$\begin{aligned}\bar{l} &= -\sum_{w \in W} \tilde{p}(w - \log_2 \tilde{p}(w) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \\ &= I(p) \qquad (58)\end{aligned}$$

The 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 $\mathcal{X}$.

The Kullback–Leibler divergence $D(p||q)$ (see Section 2.3) measures the extra average length of the codewords when an ill-adapted Shannon–Fano binary code is used, namely, a code adapted to a probability distribution q when the actual probability is p (think, for instance, of a Morse code optimised for English being 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

$$\langle l \rangle_{opt} = -\sum_{x \in \mathcal{X}} p_x \log_2 p_x,$$

giving

$$\langle l \rangle - \langle l \rangle_{opt} = D(p||q).$$

In the case of a sequence-generating source, the source X emits with probability $p(\bar{x}_N)$ a message $\bar{x}_N$ of length N written in the alphabet $\mathcal{X}$ (that is, the set of elementary events, usually called ‘symbols’). For transmission or storage purposes, the message is encoded into a binary sequence $\bar{w}$, with the additional aim of minimising the length of the coded message. A complete absence of knowledge at time t of the next symbol x_{t+1} generated by the source X is achieved in the case of a sequence of independent and equiprobable symbols (an 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 advantage of any divergence from such full randomness, which originates either from an uneven distribution of the elementary events (meaning that some symbols are observed more often), or from 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 advantage of the unevenness in 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 realisation $(x_1, \ldots, x_N)$, 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

$$\langle |\bar{w}_N| \rangle \, / \, N \leqslant H(p) + \epsilon$$

for any $N \geqslant N_\epsilon$. This means that on average, a realisation $\bar{x}_N$ of the original sequence is encoded by NH bits for N large enough.

For a correlated sequence-generating source X, compression also takes advantage of the temporal correlations, and the compression rate is controlled by the entropy rate $h(X)$, with $h(X) < H(X)$. After an optimal compression of the source X, there shoud be no further 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 about 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, which is 1. Such an optimal encoding is achieved by the 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 it follows that $h(X)$ is the average number of bits required to give an optimal encoding of a symbol emitted by the source. A realisation of the original sequence being on the average encoded by $Nh(X)$ bits, the shortest binary sequences faithfully representing the original source during a duration N will have an average length $Nh(X)$. We see here that the 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 the 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, which is also called *Kolmogorov complexity*, was introduced independently by Chaitin, Kolmogorov and Solomonoff (Chaitin 1966; Solomonoff 1978; Durand and Zvonkine 2007) to characterise 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, without reference to an ensemble of sequences or an *a priori* probability. More explicitly, a single binary 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

$$K(\bar{x}_N) \leqslant L(\bar{x}_N),$$

that is, $K(\bar{x}_N)$ provides a lower bound. Moreover, as it is related to the shortest program generating the binary sequence $\bar{x}_N$ of length N, we have

$$K(\bar{x}_N) \leqslant N + c$$

where c depends on the language (or, from another point of view, on the universal Turing machine taken as reference). Also, a specific (finite) sequence can always have a very low complexity in an *ad hoc* language, where printing this sequence is a primitive. But for arbitrarily long sequences, the language-specific contribution to $K(\bar{x}_N)$ becomes relatively negligible, hence one can speak of 'the' algorithmic complexity (Durand and Zvonkine 2007). Since entropy is not a measure of complexity, the name algorithmic complexity is misleading: characterising the complexity of a source is of little help (Gell-Mann and Lloyd 1996; Gell-Mann and Lloyd 2003). It should instead be called 'algorithmic information' as recommended in Badii and Politi (1997). The theory is of limited power for short sequences, but stationarity of the sequence is no longer mandatory. The main problem with algorithmic complexity, and its conditional extensions, is its *non-computability*. This leads us to resort to lossless compression algorithms to approximate (upper bound) algorithmic complexity – see Section 6.2.

In the same way as we can define an entropy rate h, we can 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 an n-word extracted from $\bar{x}$ (the limit does not depend on the choice of this block) (Badii and Politi 1997). Remarkably, it has been shown (Ziv and Lempel 1978) that for a stationary and ergodic source, $C(\bar{x})$ coincides up to a normalisation factor with the entropy rate of the source for almost all sequences $\bar{x}$ (typical sequences in the sense of the Shannon–McMillan–Breiman theorem), specifically,

$$C(\bar{x}) = h/\log_2 |\mathcal{X}|.$$

It follows that on 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 the Shannon entropy rate through

$$\langle K(\bar{x}_N)\rangle \sim h/\log_2 |\mathcal{X}|$$

for N large enough (Feldman and Crutchfield 1998). The notion can be fruitfully extended to the notion of the *conditional Kolmogorov complexity* $K(x|y)$ of a sequence or random object x. This 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 mentioned in Section 5.2 that two compression issues should be carefully distinguished: compression of an ensemble of sequences (source compression) and compression of a single sequence. In both cases, compression is achieved by the most efficient/economical encoding, hence coding and compression issues are thus solved jointly, and they are limited by the same bounds, involving either (Csiszár and Körner 1981; Badii and Politi 1997; Falcioni *et al.* 2003):

— the Shannon entropy rate for the compression of a source of known probability, that is, the compression of the set of sequences emitted by the source (Section 5.2); or

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

We shall focus on the second of these. Compression of a single sequence $\bar{x}$ is possible if the Kolmogorov complexity of this sequence is strictly smaller than its length: $K(\bar{x}) < |\bar{x}|$. Accordingly, a sequence $\bar{x}$ with $K(\bar{x}) = |\bar{x}|$ is said to be *incompressible.*

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

$$1 \bullet 0 \bullet 01 \bullet 10 \bullet 11 \bullet 100 \bullet 101 \bullet 00 \bullet 010 \bullet 11...$$

which delineates as a new word the shortest word that has not yet been encountered; or (Ziv and Lempel 1977)

$$1 \bullet 0 \bullet 01 \bullet 101 \bullet 1100 \bullet 1010 \bullet 001011 \bullet ...$$

which delineates as a new word the shortest subsequence that has not yet been encountered (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 us to encode the original sequence efficiently with no loss of information. Indeed, each new word appearing in the parsing is uniquely specified by the already encountered word with which it begins and the additional symbol with which it is completed. We then compute

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

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

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

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

This means that for N large enough, $\hat{L}_0(\bar{x}_N)$ not only gives 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. The Lempel–Ziv theorem also means that almost all symbolic sequences have the same compressibility features, so the computation can be performed equivalently with any typical sequence. From the Shannon–McMillan–Breiman theorem (see Section 5.1), typical sequences have a full measure, so 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 normalisation factor and provided a sufficient convergence (N enough large) is achieved (Lesne *et al.* 2009). A simpler computation is

$$\hat{L} = \frac{\mathcal{N}_w \log_2 N}{N}\tag{62}$$

with

$$\lim_{N\to\infty}\hat{L} = h$$

Replacing $\log_k$ by $\log_2$ makes the limit directly comparable to h, while the original definition is normalised 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 formalisations 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 (that is, chosing the subset of labels without involving the actual values 0 or 1 associated with each of them). Kolmogorov refined this notion into the notion of (N,ϵ)-randomness, which is relevant 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 we mean by the randomness of a sequence (Falcioni *et al.* 2003; Parisi 2003; Castiglione *et al.* 2008). An incompressible sequence $\bar{x}$, that is, one such that $K(\bar{x}) = |\bar{x}|$, is said to be *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 test of being uniformly distributed, for example, the decimal digits of π), are not algorithmically random. It was applied to real numbers by Martin-Löf (Martin-Löf 1966): by 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 the ergodic theory of dynamical systems

7.1. *Metric entropy*

Shannon entropy for discrete-valued and discrete-time stochastic processes has an exact analogue in the ergodic theory of dynamical systems. It was developed by Kolmogorov (Kolmogorov 1965) and Sinai (Sinai 1959), and is thus called metric entropy or Kolmogorov–Sinai entropy. Given a discrete-time evolution $x_{n+1} = f(x_n)$ on a compact topological space, we consider a finite partition $\mathcal{P}_0$ and the refinements generated in the course of time by the map f, namely

$$\mathcal{P}_n = \mathcal{P}_0 \vee f^{-1}(\mathcal{P}_0) \vee \ldots \vee f^{-n}(\mathcal{P}_0).$$

We then compute

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

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

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

Finally, the *metric entropy*, or *Kolmogorov–Sinai entropy* is given by (Wehrl 1978)

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

This is actually a *rate* of entropy. Note that we use ln instead of $\log_2$ in dynamical systems theory. This is a purely conventional choice, which is motivated by practical and historical reasons since the two quantities are related by a factor of ln 2, namely $h_{KS} = h \,.\, \ln 2$ (that is, $e^{h_{KS}} = 2^h$). The metric entropy h_{KS} was introduced to solve the 'isomorphism problem', that is, determining whether there is a mapping between two seemingly different dynamical systems, while preserving the dynamical and statistical relationships betwen the successive states. Since h_{KS} is invariant under any isomorphism, two dynamical systems with different values for h_{KS} are non-isomorphic.

It has also proved to be very useful in quantifying the seemingly erratic and irregular behaviour of chaotic dynamical systems (Kantz and Schreiber 1997). In some cases (for example, one-dimensional Anosov maps), there exist partitions $\mathcal{P}$, called *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 $\mathcal{P}$ (that is, the symbol labeling the corresponding element of the partition at each time) to specify uniquely the initial condition in the continuous phase space, and to reconstruct the continuous trajectory from the symbolic sequence. For a generating partition $\mathcal{P}$, we have $\widetilde{h}(\mathcal{P})$ reaches its maximum value h_{KS}, and coincides up to a factor $\ln 2$ with the Shannon entropy rate of the symbolic sequence, that is, $h_{KS} = h/\ln 2$. Accordingly, the metric entropy can also be computed on discretised trajectories. We could say more: according to the *Jewett–Krieger theorem* (Krieger 1970; Krieger 1972; Falcioni *et al.* 2003; Glasner 2003, Section 15.8), a continuous-valued dynamical system in discrete time and 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

$$e^{h_{KS}} \leqslant m < 1 + e^{h_{KS}}.$$

These results are the basis and justification for *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 to be a criterion and quantitative index of *chaos* (Laguës and Lesne 2008; Castiglione *et al.* 2008). A justification for this is the relationship between h_{KS} and the sum of positive Lyapounov exponents (Pesin 1997):

$$h_{KS} \leqslant \sum_{i,\, \gamma_i \geqslant 0} \gamma_i = \sum_i \gamma_i^+. \tag{66}$$

This inequality, which is known as the *Pesin inequality*, turns into an equality for sufficiently chaotic systems, such as Anosov systems (Ledrappier and Strelcyn 1982; Castiglione *et al.* 2008). This means that the production of information (that is, the gain in information about initial conditions from observing the trajectory for one more step) is only provided by unstable directions, so

$$h_{KS} \leqslant \sum \gamma_i^+.$$

Negative Lyapunov exponents $\gamma_i < 0$, which are associated with stable directions, play no role here. The relevance of metric entropy in data analysis for globally quantifying the temporal organisation of the evolution has been recognised in numerous applications. It is now a standard tool of non-linear time series analysis for both continuous-valued and discrete (symbolic) sequences (Kantz and Schreiber 1997).

We 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 notable point here is that ϵ-entropy can be defined and computed for any dynamical process, whether deterministic or stochastic. A behaviour

$$\lim_{\epsilon \to 0} h(\epsilon) = h_{KS}$$

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 between different kinds of stochastic processes, with trajectories becoming more irregular as the increase becomes 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 relevant when describing the overall statistical behaviour of a dynamical system. Writing $\mathcal{N}(n,\epsilon)$ to denote the maximal number of trajectories ϵ-separated over n time steps (that is, for at least one time between 0 and n, the distance between the trajectories is larger than ϵ), it is defined by

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

Unlike 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. Like h, it is defined as a rate (entropy per unit time). When a generating partition exists and allows us to investigate the statistical features of the dynamics on a reduced symbolic version, the topological entropy is given by

$$\lim_{n\to\infty}(1/n)\log\mathcal{N}_n$$

where $\mathcal{N}_n$ is the number of admissible n-words. This formula shows that topological entropy was in fact already present in Shannon's seminal paper and coincides with the notion of the *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 (see Section 4.4) allows us to unify metric and topological entropies in a unique framework: indeed, we can recover

$$h(\alpha = 1) = h_{KS}/\ln 2$$
$$h(\alpha = 0) = h_{top}/\ln 2.$$

The generalised framework of Rényi entropies is relevant in the application of the thermodynamic formalism to dynamical systems and their multifractal analysis (Badii and Politi 1997), which we shall discuss briefly in the next section.

7.3. *Thermodynamic formalism*

A unifying framework has been developed for systematically deriving all the relevant statistical features of a discrete-time dynamical system. It is formally reminiscent of the Boltzmann–Gibbs formalism in statistical mechanics, and for this reason it is called the *thermodynamic formalism* (Ruelle 1978). The basic idea is to introduce the analogue of a partition function, where the role of the n-particle 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 is given by (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 thought of as an inverse temperature. The relevant analogue of the free energy is

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

and its density is

$$J(q) = \lim_{n\to\infty} J(n,q) \tag{70}$$

where

$$J(n,q) = \frac{I(n,q)}{n}.$$

It is straightforward to show that $J(q=1)$ coincides with the metric entropy h_{KS}. Now $I(n,q)$ is just the Rényi entropy (Section 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, the 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 the following Legendre reciprocal transformations:

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

$$g(\kappa) = \inf_q [q\kappa - (q-1)J_q]. \tag{73}$$

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

7.4. *Typicality, compressibility and predictibility*

The Shannon–McMillan–Breiman 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 remaining for 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 the 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 less than ϵ from the orbit of x (Brin and Katok 1983). This means that the probability (with respect to the invariant measure μ) that two trajectories stay close together for n steps decays exponentially with n.

The relationship between the Shannon entropy rate and the metric entropy reinforces the relevance of a probabilistic description of chaotic (deterministic) dynamical systems, that is, the use of statistical descriptors to quantify their apparent randomness (Nicolis and Gaspard 1994). In particular, it gives another interpretation to their unpredictability by 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).

In addition to this global view, in which we consider a dynamical system as a random source, we could also consider each trajectory in isolation and compute the algorithmic complexity of its symbolic description. In this context, a theorem by Brudno and White (Brudno 1983; White 1993; Castiglione *et al.* 2008) says that for an autonomous ergodic dynamical system, the Kolmogorov complexity of almost all trajectories (almost all with respect to the invariant ergodic measure) is equal to h_{KS} up to a constant normalisation factor. This theorem is just the deterministic version of the Lempel–Ziv theorem (Ziv and Lempel 1978). Accordingly, for a symbolic trajectory, we have an equivalence between being unpredictable, being incompressible and being algorithmically complex (Falcioni *et al.* 2003).

8. Relation to statistical physics

8.1. *The second law of thermodynamics*

Thermodynamic entropy was 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} during an arbitrary transformation of the system from the equilibrium state A to 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 holds if and only if the transformation is reversible. For isolated systems, more precisely thermodynamically closed systems for which $\delta Q = 0$, we have $\Delta S_{th} \geqslant 0$. This statement is known as the *Second Law* 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 a molecular scale (Castiglione *et al.* 2008), and, in fact, it does not. Recent advances describe divergences from the second law arising in small systems in the form of fluctuation theorems (Gallavotti 1998; Cohen and Gallavotti 1999; Evans and Searles 2002).

The second law should not be confused with the *H-theorem* (Cercignani 1988b). The former is indeed a universal but empirical principle, which is (presumed to be) valid for any thermodynamically closed macroscopic system, while the latter is an exact (that is, rigorously proved) property of the Boltzmann kinetic equation, which is why it is called a theorem (Castiglione *et al.* 2008). This equation, which is 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 the 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 decrease in the course of time. In order to relate it to a Shannon entropy, we have 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 (Castiglione *et al.* 2008). The fact that H_B can only increase is based on the decorrelation approximation involved in the derivation of the Boltzmann equation (Cercignani 1988a). This amounts to replacing the 2-particle distributions arising in the interaction kernel by a product of one-particle distributions. Hence, the H-theorem only indirectly and approximately describes a feature of the real world, insofar as the system behaviour 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 ensembles*

The term entropy in classical statistical mechanics is basically the *Boltzmann entropy*, namely, a quantity related to the number Γ_N of N-particle microstates that 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 was proposed by Boltzmann in 1877 (Boltzmann 1877; Cercignani 1988b; Castiglione *et al.* 2008) and 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 consists of N particles).

The starting point for the statistical description is usually the microcanonical ensemble – see Castiglione *et al.* (2008) for a discussion of its relevance and validity. This corresponds to considering equiprobable microscopic configurations at a fixed volume V and fixed energy U with a tolerance δU. The Boltzmann entropy is then proportional to the logarithm of the associated phase space volume $\Gamma(N, V, U, \delta U)$. Note that δU plays no role in the thermodynamic limit $N \to \infty$. Since $\Gamma(N, V, U, \delta U)$ behaves as

$$\Gamma(N, V, U, \delta U) \sim \Gamma_1^N \delta U,$$

the Boltzmann entropy is *extensive* (proportional to N), and the contribution of δU in $\ln \Gamma$ is a higher-order term in S_B, which can be neglected for large N (that is, $\ln \delta U$ is negligible compared with $N \ln \Gamma_1$ for large N). The Shannon–McMillan–Breiman theorem allows us to make a formal bridge between the Boltzmann entropy and the 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 becomes

$$h \sim (1/N) \log_2 \mathcal{N}_N,$$

which is reminiscent of the Boltzmann entropy per particle

$$S_B/N = (k_B/N) \ln \Gamma_N.$$

The Boltzmann entropy S_B, which is defined at the level of N-particle microscopic configurations (phase space $\mathcal{X}^N$), should not be confused with the Shannon entropy of the empirical distribution (normalised histogram) of the individual states in $\mathcal{X}$ (the type $L_{\bar{x}}$ of the configuration $\bar{x}$ – see Section 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 $\mathcal{X}$, that is

$$H(L_{\bar{x}}) = -\sum_x (n_x/N) \log(n_x/N)$$

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 (in fact, it is just *postulated*) from phenomenological considerations based on the (observed) second law of thermodynamics (Clausius 1865; Gallavotti 2006). A major achievement of Boltzmann was the identification of Boltzmann entropy with thermodynamic entropy: Boltzman entropy of the macrostate (U, V, N) coincides to first order (in the thermodynamic limit $N \to \infty$) with the thermodynamic entropy $S_{th}(U, V, N)$, thereby providing Clausius entropy with a microscopic interpretation. This is justified by comparing the macroscopic predictions of statistical mechanics with the empirical laws of thermodynamics, such as in the expression of the entropy of an ideal gas. The identification requires the multiplicative factor k_B (which is equal to the ideal gas constant divided by the Avogadro number) in the definition of S_B, when compared with a dimensionless statistical entropy. A *definition* of physical entropy is only possible in the framework of quantum mechanics by exploiting its intrinsically discrete formulation. Introducing the density matrix $\widehat{D}$ characterising the quantum state of the system (a positive Hermitian operator with trace 1), the entropy is defined by

$$S(\widehat{D}) = -k_B \mathrm{Tr}(\widehat{D} \ln \widehat{D}).$$

This entropy, which was introduced by Von Neumann in 1927, measures our ignorance about the system, and accordingly vanishes in the case of a pure state, which is described by a single wave function (Wehrl 1978; Balian 2004; Balian 2005). Moreover, it is an absolute entropy, which vanishes at zero absolute temperature, in agreement with the Nernst principle. Deciding whether it is a useful physical quantity is another matter, as is determining how it can be measured and related to the macroscopic (thermodynamic) entropy. Another interpretation of this entropy is as a measurement of the amount of information gained in a quantum measurement (yielding a pure state). When considering an evolving system

$$\imath\hbar\, d\widehat{D}_t/dt = [\widehat{H}, \widehat{D}_t],$$

we have that

$$S(t) = \mathrm{Tr}(\widehat{D}_t \ln \widehat{D}_t)$$

remains constant. Any reduction of the density operator to essential variables yields a reduced operator $\widehat{D}_t^0$, for which the associated entropy $S(\widehat{D}_t^0)$ increases. See Wehrl (1978), Balian (2004) and Balian (2005) for a detailed discussion of quantum-mechanical entropy.

8.3. *The maximisation 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 with this reduced ensemble. To each value of m, there is a corresponding '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, which was derived by Einstein (Einstein 1910):

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

where

$$\Delta S(m, U) = S(U, m) - S(U) \leqslant 0$$

is proportional to the number N of particles, that is, to the size of the system. In the limit $N \to \infty$, the macrostate distribution becomes sharply peaked around the value m^*, giving the maximum Boltzmann entropy. This property follows essentially from a concentration theorem and reflects the fact that, in the thermodynamic limit $N \to \infty$, an exponentially dominant fraction of microscopic configurations are associated with the macroscopic variable m^*. For N large enough, the distribution is sharply peaked and the typical behaviour can be identified with the most probable behaviour. In other words, we observe the most probable macrostate. At leading order in N, we have that 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: for example, the fact that your coffee always cools and never draws heat from its surroundings, despite the invariance under time reversal of the microscopic dynamics (Lebowitz 1993a; Castiglione *et al.* 2008; Schulman 2010). The Liouville theorem indeed ensures the constancy in time of the density in the microcopic phase space. The answer was given by Boltzmann (Boltzmann 1877). Basically, the origin of this irreversibility lies in the non-typicality of the initial configuration when considered under the final conditions, while the final equilibrium state is typical. This asymmetry is quantified by means of the Boltzmann entropy of the two macrostates, which amounts to comparing the volumes of the phase space regions Γ_i and Γ_f associated with the prepared initial state and the final equilibrium state, respectively (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 of evolving to Γ_i rather than Γ_f are

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

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

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

In statistical mechanics textbooks (such as Chandler (1987)), the canonical ensemble is derived by imposing a fixed average energy on otherwise equiprobable microscopic configurations. However, Jaynes long ago stressed the fact that statistical mechanics can also be derived from the maximum entropy principle within a purely information-theoretic framework (Jaynes 1957a; Jaynes 1957b). As presented in a general setting in Section 3.4, this principle allows us 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 some fixed thermal energy (fixed mean square velocity, vanishing mean velocity), it yields the well-known *Maxwelll velocity distribution*:

$$\rho_N(\bar{v}_N)d^{3N}\bar{v}_N = \prod_{i=1}^{N} \rho_1(\vec{v}_i)d^3\vec{v}_i \tag{77}$$

where

$$\rho_1(\vec{v})d^3\vec{v} = e^{-mv^2/k_BT}\left(\frac{m}{2\pi k_B T}\right)^{3/2} dv_x dv_y dv_z$$

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

$$\langle mv^2/2 \rangle = 3k_B T/2. \tag{78}$$

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

$$P(\bar{x}_N|\beta) = \frac{e^{-\beta E(\bar{x}_N)}}{Z(N,\beta)} \tag{79}$$

with

$$Z(N,\beta) = \int_{\mathcal{X}^N} e^{-\beta E(\bar{x}_N)}\, d\bar{x}_N$$

where $d\bar{x}_N$ is the integration element in the $3N$-dimensional phase space (positions $\bar{x}_N$ of N particles). Note the factorisation of the distributions for the velocity and position degrees of freedom in the Maxwell and Boltzmann–Gibbs distributions, respectively, which ensures the decoupling of kinetic theory and equilibrium statistical mechanics. Compared with the microcanonical ensemble, the Boltzmann–Gibbs distribution gives different weight to the microstates defining the *canonical ensemble*. Nevertheless, the predictions of the two ensembles 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 within energy shells

$$dE = \sum x, E(x) \in [E, E + dE]$$

amounts to using the microcanonical weight

$$e^{S_B(E,N)/k_N}$$

of each shell, so we get the distribution

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

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

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

which exploits 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 us to identify

$$F = -(1/\beta) \ln Z(\beta, N)$$

with the *free energy*, the multiplier β with the inverse temperature $1/k_B T$ and the Boltzmann entropy at the maximum $E^* \equiv U$ with the thermodynamic entropy through the relation $F = U - TS_{th}$.

The maximum entropy principle could also be applied to infer the distribution of energy levels at fixed average energy, yielding

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

However, there is a caveat: the energy levels have to be discrete and non-degenerate. Indeed, the application of the 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. Here we see again that the maximum entropy principle, and more generally the Shannon entropy, is well defined and can only be used safely in discrete spaces of states (see Section 3.4). As mentioned in Section 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 approximate way (Wehrl 1978; Balian 2004).

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

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

where $\bar{x}_N$ is the system position in the full microscopic phase space (a configuration with $6N$ degrees of freedom for both the positions and the velocities of the N particles of the system), and $\rho(\bar{x}_N, t)$ is the density describing the probability of the system being in this phase space. However, it has two flaws:

- — it is defined up to an additive constant (as mentioned in Section 2.6, the continuous extension of Shannon entropy is not invariant with respect to a cooordinate change); and
- — the Liouville theorem for the microscopic dynamics ensures that $S_G(t)$ remains constant in time, even in conditions where the Second Law of thermodynamics predicts an increase in the thermodynamic entropy.

Accordingly, the Gibbs entropy in this form cannot be identified with thermodynamic entropy. Both flaws are cured by considering a coarse-grained version of Gibbs entropy (Castiglione *et al.* 2008), that is, the 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 (see Section 7.1) of the microscopic dynamics – see Castiglione *et al.* (2008) for a detailed discussion of the connections between statistical mechanics and chaos theory (it occupies several chapters and goes well beyond the scope of the present review).

8.5. *Dissipative structures and the minimum entropy production principle*

Prigogine (Prigogine 1967; Nicolis and Prigogine 1977) developed the notion of a *dissipative structure*, though examples, such as the Bénard cell, had been observed and studied well before his work. A dissipative structure is an organised pattern arising in open systems, in which local order appears at the expense of energy or matter input. Thermodynamic entropy S_{th} is only defined for equilibrium states, but for non-equilibrium states, we can define an *entropy production rate*. The entropy production can be decomposed into

$$dS_{th} = dS_{irr} + dS_{exch}$$

where dS_{exch} is the contribution due to exchanges of matter and energy. At steady state, $dS_{th} = 0$, but we can have $dS_{irr} > 0$ at the expense of $dS_{exch} < 0$, which is precisely the case for dissipative structures. dS_{irr} is often thought of as a measure of the irreversibility of the system, but its definition and interpretation are restricted to the framework of the thermodynamics of irreversible processes, and is itself 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):

$$d[(dS_{th}/dt)_{irr}] = 0$$

where $(dS_{th}/dt)_{irr}$ is the entropy production rate due to irreversible processes (Prigogine 1967). Nevertheless, this principle, expressing the stability of a non-equilibrium steady state, can only be rigorously derived under very restrictive conditions (assumptions of local equilibrum thermodynamics and linear response, isotropic medium, time independence of boundary conditions and linear response coefficients, and isothermal system in mechanical and thermal equilibrium). Its general validity and application are thus highly questionable (Kay 1984). As emphasised in Mahara and Yamaguchi (2010), while entropy production could be used to discriminate between different patterns, minimising entropy production is not a valid criterion for pattern selection.

Thirty years ago, Jaynes produced a very deep and stimulating analysis (Jaynes 1980) in which he pointed out that Kirchhoff laws for determining the distribution of currents in an electric circuit are already fully determined by conservation laws, with no need for an additional entropic criterion. He raised this question in the context of the non-equilibrium extension of Gibbs work on the characterisation of heterogeneous equilibrium (phase coexistence) using a variational principle on thermodynamic entropy. It should be stressed at this point that all the derivations and justifications of the minimum entropy production 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. *Non-equilibrium systems and the chaotic hypothesis*

Going beyond linear response theory, if we are to give a general definition of entropy and entropy production in far-from-equilibrium systems, we will need 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 a hyperbolic dynamical system (Cohen and Gallavotti 1999; Evans and Searles 2002; Gallavotti 2006). This so-called *chaotic hypothesis* is the far-from-equilibrum analogue of the assumption of ergodicity or molecular chaos (the 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 weights given by the non-equilibrium steady state measure $\rho(dx)$, that is, $\int e(x)\rho(dx)$ (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

$$H_n^R = -\sum_{\bar{w}} p_n(\bar{w})\log_2 p_n(\bar{w}^R),$$

and

$$\bar{w}^R = (w_n, \ldots, w_1).$$

He then showed that for Markov processes (at least), the entropy production rate is given by

$$dS/dt = h^R - h \tag{85}$$

where

$$S(t) = -\sum_{w} p_t(w)\log_2 p_t(w).$$

This time-reversal symmetry breaking, which is reflected in entropy production, corresponds to the fact that the distributions of the incoming and outgoing particles differ strongly. The latter is finely correlated due to the interactions between the particles within the system. Observing the time-reversed steady state would require us to prepare the incoming flux of particles according to such an intricately correlated distribution. The formula (85) provides us with a rigorous and quantitative expression of the relationship between irreversibility and the entropy production rate in a non-equilibrium stationary state. We stress that the irreversibility of a system driven far-from-equilibrium by fluxes is fundamentally different from the irreversibility observed in the relaxation of an isolated system after lifting a constraint, as discussed in Section 8.3.

8.7. *Thermodynamic cost of computation*

In addition to the formal link based on the maximum entropy inference of the Boltzmann–Gibbs distribution (Jaynes 1957a; Jaynes 1957b), another relationship between statistical mechanics and information theory is the paradox called Maxwell's demon, which 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 admits particles from the cold compartment if their velocity is greater than

$$\sqrt{3k_BT_2/m},$$

and lets particles leave the hot compartment if their velocity is less 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 law. Brillouin (Brilllouin 1951a; Brillouin 1951b) suggested a way out of the Maxwell's demon paradox by showing, in a specific example, that work has to be performed in order to achieve a measurement. In other words, the demon needs information about the particle velocity, which has a cost (Brillouin 1956). In a simpler variant, the compartments are at the same temperature and the demon allows particles to pass from the first compartment to the second but prevents them going in the opposite direction, so that in the final state of the system, all N particles will be in the second compartment. The decrease in the thermodynamic entropy by $k_B N \ln 2$ is 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 to 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 the general inequality,

$$W_{meas} + W_{eras} \geqslant k_B T\ I$$

where W_{meas} is the work cost of making the measurement, W_{eras} is the work cost of erasing the memory storing the result and I is the mutual information shared between the measured system and the memory (that is, the information gained about the system in the measurement). The work W_{meas} could vanish in some instances, in which case Landauer's result is recovered, while the complete inequality is also consistent with Brillouin's result.

9. Typicality and statistical laws of collective behaviour

9.1. *Probabilistic modelling and subjective probabilities*

The notion of statistical entropy, being relative to a probability distribution, leads us 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.

The *frequentist* and *subjective* (or Bayesian) viewpoints are well-known alternatives for considering the reconstruction and epistemic status of a probability distribution (Jaynes 1957a; Jaynes 1957b; Bricmont 1995). Both viewpoints yield efficient reconstruction methods. The frequentist viewpoint belongs to the realm of statistical estimation from independent samples, and is essentially based on the law of large numbers (Samengo 2002). The Bayesian viewpoint belongs to the realm of learning and recursive algorithms, with the updating from new data of a prior distribution into a posterior one. A seminal paper by Cox (Cox 1946) underlined the fact that the frequentist definition is inseparable from the existence of an ensemble (at least conceptually). He called the Bayesian viewpoint the idea of 'reasonable expectation', which is related to the notion of 'degree of rational belief' formulated by Keynes. Some Bayesian probabilities cannot be cast in an ensemble (that is, 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 might prefer to think of the probabilistic proof of the existence of Santa Claus given by Howard Buten (Buten 1989).

Jaynes (Jaynes 1973; Jaynes 1982b) had already highlighted the choice between the frequentist view, trying to estimate the frequencies of 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, in the form of the maximum entropy principle (Jaynes 1982a), for setting up probability distributions on the basis of partial knowledge, which was discussed in Jaynes (1957a) and Jaynes (1957b) in the context of statistical physics.

The subjective (Bayesian) view of probabilities (Cox 1946; de Finetti 1970; Gillies 2000; Balian 2005) encapsulates *a priori* but incomplete knowledge, such as a set of possible states, but also apparent randomness at the observation scales. In both cases, this means that our limited perception is best represented by a probability distribution, irrespective of whether the nature of the system is stochastic or not. The probabilistic aspect of the description lies only in our representation of the reality, with no intention of saying anything about the nature of the real phenomenon. A probability distribution does not aim to be an intrinsic and absolute character ruling the system behaviour (as it is in quantum mechanics), but only the best operational and faithful account of *our* knowledge of the system. See Jaynes (1957a), Jaynes (1957b) and Jaynes (1973) for a detailed and substantiated discussion of this viewpoint. Such a pragmatic view of a probabilistic description is currently adopted for chaotic dynamic systems (Nicolis and Gaspard 1994), where we give up a description in terms of the deterministic trajectories in favour of a stationary and global description in terms of an invariant measure (the latter is just the distribution describing the probability of occupying the phase space). In any case, probability theory can be thought of purely as an operational tool, even if there is no stochasticity involved in the problem, as in the probabilistic formulation of some properties in number theory (Cox 1946; Ford 2007). The 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.

Note that in the realm of *classical* physics, we cannot assess whether a system is intrinsically probabilistic unless we start at the quantum level (Lesne 2007). However, the observed randomness of coin tossing, say, has a very different nature to quantum uncertainty. For the most part, it can be accounted for by arguing about the chaotic nature of the coin motion when it is flipped in the air. The randomness thus originates in our lack of knowledge of the initial conditions and countless minute influences experienced by the coin while it is tossed. The probabilistic nature is not that of the system, but belongs to one of our possible descriptions. In particular, it depends essentially on the scale of the description. A well-known example is diffusion, for which a hierarchy of descriptions exist depending on the scale and the level of coarse graining, ranging from a deterministic reversible description (molecular dynamics), through several stochastic descriptions (the master equation, random walks and the Fokker–Planck equation, the Langevin equation) to a deterministic irreversible description (the Fick law and the century-old diffusion equation) (Castiglione *et al.* 2008).

Finally, we shall now consider 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, explicitly distinguishing subpopulations α with fraction f_α (hence $\sum_\alpha f_\alpha = 1$) allows us 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 with a choice between a detailed description using an array $[(p_\alpha)_\alpha]$ and a global probabilistic view in the form of an effective description of the knowledge available at the scale of the population given by 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, which prevents any further attempt to talk about any would-be ‘intrisic stochastic nature’ of a system. We deal only with models, that is, 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 behaviours in physics*

We have just argued that probability, in the subjective viewpoint, is a privileged framework allowing us to take into account in a unified way observation scales and the limits they set on our perceptions and representations. It is also a unified framework for investigating collective behaviours and unravelling the mathematical structures underlying emergent properties. A central physical example of this is thermodynamic behaviour. This behaviour corresponds to a sharp distribution for macroscopic quantities, meaning that almost all microscopic configurations yield the same macroscopic values. In such a case, the probability distribution of the microscopic configurations (that is, their respective frequencies of occurrence) has almost no macroscopic consequences when it is non-singular. Accordingly, thermodynamics relies almost entirely on universal statistical laws, mainly the law of large numbers and the central limit theorem. The counterpart of this universality is that macroscopic behaviour is quite insensitive to microscopic features. In particular, knowing the macroscopic behaviour gives no insight into the microscopic distribution and is useless for inferring any knowledge about the microscopic elements. The success of the maximum entropy approach provides evidence for the fact that thermodynamic laws are based on universal statistical laws governing the structure and features of emergent behaviours, rather than on specific physical laws (Jaynes 1957a; Jaynes 1957b). More generally, statistical laws express rules of collective behaviour, no matter what the physical nature of the elements and their interactions may be. They state a general *mathematical* property of *any* high-dimensional system (for example, many-body systems in physics or long messages in communication theory). They account, for instance, for the ubiquitousness of Gaussian distributions (resulting from the central limit theorem).

The same all-or-nothing law arises in different contexts and under different names (see Sections 3.1 and Section 5.1):

- the law of large numbers and Lévy's *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);
- the *asymptotic equipartition property* (Shannon 1948; Cover and Thomas 2006) in information theory;
- the *ergodic theorem* in dynamical systems.

Close connections can be established between these different laws (Lesne 1998). They can be viewed as a universal mathematical structure of collective behaviours.

Let us consider again the Shannon–McMillan–Breiman theorem – see Section 5.1. The property

$$\lim_{n\to\infty}(1/n)\log_2 \hat{P}_n - h = 0$$

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. The 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), and also known as the *Hewitt–Savage 0-1 law*, which 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, while for any $h' \neq h$,

$$\lim_{n\to\infty}(1/n)\log_2 \hat{P}_n(\bar{x}) - h' = 0$$

has a null probability of being true.

The predictability and simplicity of macroscopic physical phenomena arise from the fact that at the macroscopic level, a wealth of behaviours result from a bottom-up integration and emergence. They are governed 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 only involved in prescribing the *universality class* of the emergent behaviour. Basically, we have to discriminate between systems with short-range correlations, which display scale separation between microscopic and macroscopic levels, and systems with long-range correlations, which are associated with criticality and anomalous statistical laws (Lesne 1998; Castiglione *et al.* 2008). A typical example is *diffusion*, which passes from normal to anomalous in the case of long-range correlations (self-avoiding walks). This corresponds to the passage from the universality class of the Wiener process to that of fractal Brownian motions. Another anomaly is observed in diffusive behaviour when the variance of the elementary steps diverge, corresponding to the passage from the central limit theorem assessing convergence to a Gaussian distribution to generalised limit theorems assessing convergence to Lévy stable laws (Lesne 1998; Castiglione *et al.* 2008). In general, universality and robustness arise in physics when the statistics and geometry are sufficient to determine 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 the system belongs to.

9.3. *Typicality*

The following list describes several notions of typicality, some of which have already been discussed earlier in the paper:

(1) A notion based on concentration theorems for a configuration or a sequence $(X_i)_i$ of independent and identical elements (see Section 3.1):
When reasoning about the configuration or sequence type, typical sequences belong to the most populated type. 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 with the most populated one. The law of large numbers can be thought of as a statement about the typical behaviour of the empirical average

$$\widehat{m}_N = (1/N)\sum_{i}^{N} X_i.$$

In other words, 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 realisations of the sequence satisfying $|\widehat{m}_N - m] < \epsilon$ is smaller than δ, meaning that, asymptotically, almost all realisations of the sequence are typical with respect to the behaviour of the empirical average.

(2) A notion based on the Sanov theorem for sequences of independent and identical elements (see Section 3.2):
A pair of sequences $(\bar{x}_N, \bar{y}_N)$ is *jointly typical* if each individual sequence is typical with respect to h_X and h_Y, respectively, and if

$$| - (1/N)\log_2 P_N(\bar{x}_N, \bar{y}_N) - h_{X,Y}|$$

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 generalised asymptotic equipartition property:
Specifically, 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 the Shannon–McMillan–Breiman theorem for correlated sequences (see Section 5.1):
A sequence $\bar{x}_N$ of length N generated by a source of entropy rate h is typical if

$$|-(1/N)\log_2 P_N(\bar{x}_N) - h|$$

is small. The realisations with probabilities that satisfy the Shannon–McMillan–Breiman estimate form the typical set (strictly speaking, this is only defined once some tolerance ϵ is given), which is quite small but of probability close to 1. For correlated binary sequences of length N, we have 2^N possible realisations, but only about 2^{Nh} typical ones.

(5) A notion of typicality connected with the *Birkhoff's ergodic theorem*:
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 if any f-invariant subset has 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, with $\mu(\mathcal{X} - \mathcal{X}_\phi) = 0$) such that for any $x \in \mathcal{X}_\phi$, we have

$$\lim_{N\to\infty} (1/N)\sum_{i=0}^{N-1} \phi[f^i(x)] = \int_{\mathcal{X}} \phi(x)d\mu(x).$$

In this sense, the elements of $\mathcal{X}_\phi$ are typical since their behaviours are all identical and coincide with an average quantity, that is, the time averages along a typical trajectory equal the ensemble averages.
In Section 6.2, we encountered another ergodic theorem (Ziv and Lempel 1978), which endows 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 normalisation factor.

The converse of typicality is rarity. Exceptional events are non-typical events. However, several notions overlap, and should be carefully distinguished. The sequence 123456789 can be called 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 we implicitly compare the number of sequences $(n, n+1, \ldots, n+8)$ with the set of sequences that are not of this form. In other words, we compare the types of the sequences rather than the sequences themselves. This yields two ways of being random: either having the largest algorithmic complexity or belonging to the most represented type. 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 a typical sequence. This is exactly the meaning of algorithmic complexity. From this viewpoint, typicality coincides with (full) randomness. Note that in all cases, typicality is an asymptotic feature, and is well defined only in the limit

$N \to \infty$ where N is the sequence length or number of elements. Genericity and typicality arguments are ubiquitous in statistical physics, but we suspect that they cannot be applied blindly in biology, where rare events could play an essential role – see Section 9.5.

9.4. *Entropy, order and disorder*

In this section we shall discuss in what way statistical entropy can be thought of as a *measure of disorder*. However, although currently widespread and superficially appealing, this view is flawed, and also somewhat fuzzy since it requires us first to define what we mean by disorder, beyond the plain and non-technical meaning of the word. We can again take two viewpoints, corresponding to the choice between the (statistical) information-theoretic approach and the algorithmic one.

The first viewpoint is that order (for example, 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 since it can be 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 a homogeneous/stationary distribution that is invariant with respect to any translation. Specifying a structure amounts to specifying a lack of invariance. This corresponds to a decrease in the entropy compared with the fully random case (Leyton 2001).

Another viewpoint is that when we speak of order and disorder, we are, effectively, comparing sets. For instance, the sequence 123456 is ordered insofar as it is a representative of the set $\{(n, n+1, \ldots, n+5)\}$, as opposed to its complement, in other words, any sequence that is not of the form $(n, n+1, \ldots, n+5)$. The difficulty with this view is that it requires a prior and necessarily subjective delineation of an ensemble from a single observation. Order and disorder are then relative to the mind perceiving them.

Gorban gives the very inspiring example of a castle, a garden of stones and any pile of stones (Gorban 2007): the relevant comparison for the castle is with any pile of stones that is not a castle, but for the gardener, the garden of stones also has to be compared with any pile of stones that is not that garden of stones. As such, a garden of stones is as ordered and non-typical as a castle, but it is less ordered when using the criterion of individual configuration complexity. A garden of stones is viewed very differently by the gardener and most other people. In a formalised way, the relevant entropy is that of the coarse-grained 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 or patterns, so the class of configurations looking like the given one is very large; the given configuration is then said to be disordered. Rather than a measure of disorder, entropy is a measure of the typicality of the disorder, that is, a measure of degeneracy: how many configurations share the same macroscopic observables and constraints.

In the framework of equilibrium statistical mechanics, the key idea in describing order is to consider an order parameter m. This is a macroscopic feature $m(\bar{x}_N)$ measuring the overall organisation in a physically or intuitively relevant way, such as the magnetisation of ferromagnetic materials, the mass density in pure liquids or the orientation of molecules in liquid crystals. The same reasoning as we used in equations (80)–(82) in Section 8.4 can be followed, but now considering the order parameter m instead of total energy E. The contribution of the phase space volume with fixed m was given in equation (75) in Section 8.3 in terms of the Boltzmann entropy $S(m)$. We can thus obtain the distribution of the (macroscopic) order parameter, involving the free energy

$$F(m) = U - TS(m),$$

as

$$P(m) \sim e^{-\beta F(m)}. \tag{86}$$

This distribution depends on the temperature, and, in particular, its behaviour as temperature varies reveals thermal phase transitions that are associated with a qualitative change in the macroscopic order displayed by the system (Chandler 1987).

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

It is now acknowledged that the basic role of food is to provide enough energy to the organism for it to free itself from the entropy produced while it is alive. In this respect, a living organism is an instance of a dissipative structure – see Section 8.5. It is essential here that the balance is written in terms of free energy. Boltzmann pointed out that life is a struggle for entropy[†]. Schrödinger expanded on this view with the concept of negative entropy, which is the opposite of an entropy (Schrödinger 1944), or *negentropy*, a term first coined by Brillouin (Brillouin 1953). The problem with such formulations is that the entropy of a driven system (an open system driven far from equilibrium by fluxes) is undefined (Ruelle 2003), and the second law, which Schrodinger's statement implicitly refers to, does not apply directly to open systems, and thus, in particular, not to living systems. So this idea should only be taken as giving an intuitive understanding, and not as a technical or constructive theory.

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

Finally, the universal statistical laws (see Section 9.2) underlying thermodynamic behaviour are valid in physical systems under the quite mild condition that correlations between elements are summable. They only fail at critical points, where they are replaced by self-similar features (Laguës and Lesne 2008; Castiglione *et al.* 2008). By contrast, their validity is highly questionable in complex systems, and in particular for living

[†] The specific quote is '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.' (Boltzmann 1877; Cercignani 1988b)

systems, because of top-down causation. Here we mean the existence of feedback loops through which collective behaviours and emergent features can influence not only the elementary states, but also their rules of interaction and evolution. Such feedbacks from the macroscopic level to the underlying levels prevent the application of the law of large numbers and the central limit theorem. At the moment, information theory is only bottom-up, and is not suited to taking into account how an emerging feature modifies the state space or the rules of an element. A first direction for extending this would be to change the description level and investigate the relations between the distribution of probabilities to capture invariants and predictable facts (Lesne and Benecke 2008). Another direction would be to focus on interlevel relations and consistency, in the hope of finding some universality in the regulatory schemes that is absent when we restrict attention to a single level of organisation. In any case, novel statistical laws involving the reciprocal coupling and consistency between the different levels of organisation at their heart will need to be developed.

Acknowledgements

This paper has benefited from my discussions with Pierre Cartier, Giuseppe Longo and Mioara Mugur-Schächter.

References

Algoet, P. H. and Cover, T. M. (1988) A sandwich proof of the Shannon–McMillan–Breiman theorem. *Annals of Probability* **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.

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.

Balian, R. (2004) Entropy, a protean concept. In: Dalibard, J., Duplantier, B. and Rivasseau, V. (eds.) *Entropy*, Poincaré Seminar 2003, Birkhaüser 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. *Journal of Physics: Condensed Matter* **22** 063101.

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

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

Boltzmann, L. (1877) Über die Beziehung zwisschen dem zweiten Haubtsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive dem Sätzen über das Wärmegleichgewicht. ('On the Relation between the Second Law of the Mechanical Theory of Heat and the Probability Calculus with respect to the Propositions about Heat-Equivalence'.)

Wiener Berichte **76** 373–435. (Included in *Wissenschaftliche Abhandlungen* **2**, paper 42 (1909) Barth, Leipzig; reissued in 1969, Chelsea, New York.)

Breiman, L. (1957) The individual ergodic theorem of information theory. *Annals of Mathematical Statistics* **28** 809–811. (Correction: (1957) **31** 809–810.)

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

Brillouin, L. (1951a) Maxwell's demon cannot operate: Information and entropy. *Journal of Applied Physics* **22** 334–337.

Brillouin, L. (1951b) Physical entropy and information. *Journal of Applied Physics* **22** 338–343.

Brillouin, L. (1953) Negentropy principle of information. *Journal of Applied Physics* **24** 1152–1163.

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

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

Brudno, A.A. (1983) Entropy and the complexity of the trajectory of a dynamical system. *Transactions of the Moscow Mathematical Society* **44** 127–152.

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

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

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-Verlag.

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. *Journal of the 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. *Journal of Statistical Physics* **96** 1343–1349.

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

Cox, R.T. (1946) Probability, frequency, and reasonable expectation. *American Journal of Physics* **14** 1–13.

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

Csiszár, I. (1998) The Method of types. *IEEE Transactions on Information Theory* **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.

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

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

Einstein, A. (1910) Theorie der Opaleszenz von homogenen Flüssigkeiten und Flüssigkeitsgemischen in der Nähe des kritischen Zustandes. *Annalen der Physik (Leipzig)* **33** 1275–1298. (English translation: Theory of opalescence of homogeneous liquids and mixtures of liquids in the vicinity of the critical state. In: Alexander, J. (ed.) *Colloid Chemistry*, Rheinhold, 1913, Volume I, 323–329. Reprinted in: Stachel, J. (1987) (ed.) *The Collected Papers of Albert Einstein*, Princeton University Press **3** 231–249.)

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

Evans, D. J. and Searles, D. J. (2002) The fluctuation theorem. *Advances in Physics* **51** 1529–1585.

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

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

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

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

Frank, S. A. (2009) The common patterns of nature. *Journal of Evolutionary Biology* **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. *Journal of Statistical Physics* **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: Gell-Mann, M. and Tsallis, C. (eds.) *Nonextensive Entropy – Interdisciplinary Applications*, Oxford University Press 387–398.

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

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

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

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. *International Journal of Theoretical Physics* **25** 907–938.

Gray, R. M. (1990) *Entropy and information theory*, Springer. (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. *Journal of Statistical Physics* **117** 739–772.

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

Honerkamp, J. (1998) *Statistical physics*, Springer-Verlag.

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

Jaynes, E. T. (1957a) Information theory and statistical mechanics Part I. *Physical Review* **106** 620–630.

Jaynes, E. T. (1957b) Information theory and statistical mechanics Part II. *Physical Review* **108** 171–190.

Jaynes, E. T. (1973) The well-posed problem. *Foundations of Physics* **3** 477–493.

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

Jaynes, E. T. (1980) The minimum entropy production principle. *Annual Review of Physical Chemistry* **31** 579–601.

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

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

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

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.

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

Kolmogorov, A. N. (1965) Three approaches to the quantitative definition of information. *Problems of Information Transmission* **1** 1–7.

Krieger, W. (1970) On entropy and generators of measure-preserving transformations. *Transactions of the American Mathematical Society* **149** 453–464.

Krieger, W. (1972) On unique ergodicity. In: *Proceedings Sixth Berkeley Symposium* **2**, University of California Press 327–346.

Kullback, S. and Leibler, R. (1951) On information and sufficiency. *Annals of Mathematical Statistics* **22** 79–86.

Laguës, M. and Lesne, A. (2008) *Invariances d'échelle*, 2nd edition, Belin, Paris. (English translation (2011) *Scaling*, Springer-Verlag.)

Landauer, R. (1961) Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development* **5** 183–191.

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

Lebowitz, J. L. (1993b) 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. *Ergodic Theory and Dynamical Systems* **2** 203–219.

Lempel, A. and Ziv, J. (1976) On the complexity of finite sequences. *IEEE Transactions on Information Theory* **22** 75–81.

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

Lesne A. (2007) Discrete *vs* continuous controversy in physics. *Mathematical Structures in Computer Science* **17** 185–223.

Lesne, A. and Benecke, A. (2008) Feature context-dependency and complexity reduction in probability landscapes for integrative genomics. *Theoretical Biology and Medical Modelling* **5** 21.

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

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

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.

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. *Information and Control* **9** 602–619.

McMillan, B. (1953) The basic theorems of information theory. *Annals of Mathematical Statistics* **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 and Fractals* **4** 41–57.

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

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

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

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

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

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

Rached, Z., Alajaji, F. and Campbell, L. (2001) Rényi's divergence and entropy rates for finite alphabet Markov sources. *IEEE Transactions on Information Theory* **47** 1553–1562.

Robert, C. (1990) An entropy concentration theorem: applications in artificial intelligence and descriptive statistics. *Journal of Applied Probability* **27** 303–313.

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

Ruelle, D. P. (2003) Extending the definition of entropy to nonequilibrium steady states. *Proceedings of the National Academy of Sciences of the United States of America* **100** 3054–3058.

Samengo I. (2002) Estimating probabilities from experimental frequencies. *Physical Review E* **65** 046124.

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

Sagawa, T. and Ueda, M. (2009) Minimal energy cost for thermodynamic information processing: measurement and information erasure. *Physical Review Letters* **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 Technical Journal* **27** 379–423.

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

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

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

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

Solomonoff, R. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory* **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). *Zeitschrift für Physik* **53** 840–856.

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

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

Van Campenhout, J. M. and Cover, T. M. (1981) Maximum entropy and conditional entropy. *IEEE Transactions on Information Theory* **27** 483–489.

Vovk, V. and Shafer, G. (2003) Kolmogorov's contributions to the foundations of probability. *Problems of Information Transmission* **39** 21–31.

Werhl, A. (1978) General properties of entropy. *Reviews of Modern Physics* **50** 221–261.

White, H. (1993) Algorithmic complexity of points in dynamical systems. *Ergodic Theory and Dynamical Systems* **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 Transactions on Information Theory* **35** 1250–1258.

Ziv, J. and Lempel, A. (1977) A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* **23** 337–343.

Ziv, J. and Lempel, A. (1978) Compression of individual sequences by variable rate coding. *IEEE Transactions on Information Theory* **24** 530–536.

Zuk, O., Kanter, I. and Domany, E. (2005) The entropy of a binary hidden Markov process. *Journal of Statistical Physics* **121** 343–360. (Conference version: Aymptotics of the entropy rate for a hidden Markov process. *Proceedings DCC'05* 173–182.)

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