Roman Frigg and Charlotte Werndl

Entropy: a guide for the perplexed

Book section

Entropy – A Guide for the Perplexed

Roman Frigg and Charlotte Werndl*

June 2010

Contents

1 Introduction

Entropy is ubiquitous in physics, and it plays important roles in numerous other disciplines ranging from logic and statistics to biology and economics. However, a closer look reveals a complicated picture: entropy is defined differently in different contexts, and even within the same domain different notions of entropy are at work. Some of these are defined in terms of probabilities, others are not. The aim of this chapter is to arrive at an understanding of some of the most important notions of entropy and to clarify the relations between them. In particular, we discuss the question what kind of probabilities are involved whenever entropy is defined in terms of probabilities:

*The authors are listed alphabetically; the paper is fully collaborative. To contact the authors write to r.p.frigg@lse.ac.uk and charlotte.werndl@queens.ox.ac.uk.

are the probabilities chances (i.e., physical probabilities) or credences (i.e., degrees of belief)?

After setting the stage by introducing the thermodynamic entropy (Section 2), we discuss notions of entropy in information theory (Section 3), statistical mechanics (Section 4), dynamical systems theory (Section 5) and fractal geometry (Section 6). Omissions are inevitable; in particular, space constraints prevent us from discussing entropy in quantum mechanics and cosmology.[1]

2 Entropy in Thermodynamics

Entropy made its first appearance in the middle of the 19th century in the context of thermodynamics (TD). TD describes processes like the exchange of heat between two bodies or the spreading of gases in terms of macroscopic variables like temperature, pressure and volume. The centre piece of TD is the so-called Second Law of TD, which, roughly speaking, restricts the class of physically allowable processes in isolated systems to those that are not entropy decreasing. In this section we introduce the TD entropy and the Second Law.[2] We keep this presentation short because the TD entropy is not a probabilistic notion and therefore falls, strictly speaking, outside the scope of this book.

The thermodynamic state of a system is characterised by the values of its thermodynamic variables; a state is an *equilibrium state* if, and only if (iff), all variables have well-defined and constant values. For instance, the state of a gas is specified by the values of temperature, pressure and volume, and the gas is in equilibrium if these have well-defined values which do not change over time. Consider two states A and B. A process that changes the state of the system from A to B is *quasistatic* iff it only passes through equilibrium states (i.e., if all intermediate states between A and B are also equilibrium states). A process is *reversible* iff it can be exactly reversed by an infinitesimal change in the external conditions. If we consider a cyclical process – a process in which the beginning and the end state are the same – a reversible process leaves the system *and its surroundings* unchanged.

The Second Law (in Kelvin's formulation) says that it is impossible to devise an engine which, working in a cycle, produces no effect other than the extraction of heat from a reservoir and the performance of an equal amount

[1] Hemmo & Shenker (2006) and Sorkin (2005) provide good introductions to quantum and cosmological entropies, respectively.

[2] Our presentation follows Pippard (1966, pp. 19-23, 29-37). There are also many different (and non-equivalent) formulations of the Second Law (see Uffink 2001).

of mechanical work. It can be shown that this formulation implies that

$$\oint \frac{dQ}{T} \leq 0, \tag{1}$$

where dQ is the amount of heat put into the system and T the system's temperature. This is known as *Clausius' Inequality*. If the cycle is reversible, then the inequality becomes an equality. Trivially, this implies that for reversible cycles

$$\int_A^B \frac{dQ}{T} = -\int_B^A \frac{dQ}{T} \tag{2}$$

for *any* paths from A to B and from B to A, and the value of the integrals only depends on the beginning and the end point.

We are now in a position to introduce the thermodynamic entropy S_{TD}. The leading idea is that the integral in euqation (2) gives the entropy difference between A and B. We can then assign an absolute entropy value to every state of the system by choosing one particular state (we can choose any state we please!) as the reference point, choosing a value for its entropy $S_{TD}(A)$, and then define the entropy of all other points by

$$S_{TD}(B) := S_{TD}(A) + \int_A^B \frac{dQ}{T}, \tag{3}$$

where the change of state from A to B is *reversible*.

What follows from these considerations about irreversible changes? Consider the following scenario: we first change the state of the system from A to B on a quasi-static irreversible path, and then go back from B to A on a quasi-static reversible path. It follows from equations (1) and (3) that

$$S_{TD}(B) - S_{TD}(A) \leq \int_A^B \frac{dQ}{T}. \tag{4}$$

If we now restrict attention to *adiathermal* processes (i.e., ones in which temperature is constant), the integral in euqation (4) becomes zero and we have

$$S_{TD}(B) \geq S_{TD}(A). \tag{5}$$

This is often referred to as the Second Law, but it is important to point out that it is only a special version of it which holds for adiathermal processes.

S_{TD} has no intuitive interpretation as a measure of disorder, disorganisation or randomness (as often claimed). In fact such considerations have no place in TD.

We now turn to a discussion of the information theoretic entropy, which, unlike the S_{TD} is a probabilistic concept. At first sight the information

theoretic and the thermodynamic entropy have noting to do with each other. This impression will be dissolved in Section 4 when a connection is established via the Gibbs entropy.

3 Information Theory

Consider the following situation (Shannon 1949). There is a source S producing messages which are communicated to a receiver R. The receiver registers them, for instance, on a paper tape.[3] The messages are discrete and sent by the source one after the other. Let $m = \{m_1, ..., m_n\}$ be a complete set of messages (in the sense that the source cannot send messages other than the m_i). The production of one message is referred to as a *step*.

When receiving a message, we gain information, and depending on the message, more or less information. According to Shannon's theory, information and uncertainty are two sides of the same coin: the more uncertainty there is, the more information we gain by removing the uncertainty.

Shannon's basic idea was to characterise the amount of information gained from the receipt of a message as a function which depends only on how likely the messages are. Formally, for $n \in \mathbb{N}$ let V_m be the set of all probability distributions $P = (p_1, \ldots, p_n) := (p(m_1), \ldots, p(m_n))$ on $m_1, \ldots, m_n$ (i.e., $p_i \geq 0$ and $p_1 + \ldots + p_n = 1$). A reasonable measure of information is a function $S_{S,d}(P) : V_m \to \mathbb{R}$ which satisfies the following axioms (cf. Klir 2006, section 3.2.2):

1. *Continuity.* $S_{S,d}(p_1, \ldots, p_n)$ is continuous in all its arguments $p_1, \ldots, p_n$.

2. *Additivity.* The information gained of two independent experiments is the sum of the information of the experiments, i.e., for $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_k)$, $S_{S,d}(p_1q_1, p_1q_2, \ldots, p_nq_k) = S_{S,d}(P) + S_{S,d}(Q)$.

3. *Monotonicity.* For uniform distributions the information increases with n. That is, for any $P = (\frac{1}{n}, \ldots, \frac{1}{n})$ and $Q = (\frac{1}{k}, \ldots, \frac{1}{k})$, for arbitrary $k, n \in \mathbb{N}$ we have: if $k > n$, then $S_{S,d}(Q) > S_{S,d}(P)$.

4. *Branching.* The measure of information is independent of how the process is divided into parts. That is, for $(p_1, \ldots, p_n)$, $n \geq 3$, divide $m = \{m_1, \ldots, m_n\}$ into two blocks $A = (m_1, \ldots, m_s)$ and

[3] We assume that the channel is noiseless and deterministic, meaning that there is a one-to-one correspondence between the input and the output.

$B = (m_{s+1}, \ldots, m_n)$, and let $p_A = \sum_{i=1}^{s} p_i$ and $p_B = \sum_{i=s+1}^{n} p_i$. Then

$$S_{S,\,d}(p_1, ..., p_n) = S_{S,\,d}(p_A, p_B) + p_A S_{S,\,d}(\frac{p_1}{p_A}, ..., \frac{p_s}{p_A}) + p_B S_{S,\,d}(\frac{p_{s+1}}{p_B}, ..., \frac{p_n}{p_B}).^{4} \tag{7}$$

5. *Bit normalisation.* By convention, the average information gained for two equally likely messages is one *bit* ('binary digit'): $S_{S,\,d}(1/2, 1/2) = 1$.

There is exactly one function satisfying these axioms, the *discrete Shannon Entropy*:[5]

$$S_{S,\,d}(P) \;:=\; -\sum_{i=1}^{n} p_i \log[p_i], \tag{8}$$

where 'log' stands for the logarithm to the basis of two.[6] Any change toward equalization of $p_1, \ldots, p_n$ leads to an increase of $S_{S,\,d}$, which reaches its maximum, $\log[n]$, for $p_1 = ... = p_n = 1/n$. Furthermore, $S_{S,\,d}(P) = 0$ iff all p_i but one equal zero.

What kind of probabilities are invoked in Shannon's scenario? Approaches to probability can be divided into two broad groups.[7] First, epistemic approaches take probabilities to be measures for degrees of belief. Those who subscribe to an objective epistemic theory take probabilities to be degrees of rational belief, whereby 'rational' is understood to imply that given the same evidence, all rational agents have the same degree of belief in any proposition. This is denied by those who hold a subjective epistemic theory, regarding probabilities as subjective degrees of belief that can differ between persons even if they are presented with the same body of evidence. Second, ontic approaches take probabilities to be part of the 'furniture of the world'. The two most prominent ontic approaches are frequentism and a propensity view. On the frequentist approach, probabilities are long run frequencies of certain events. On the propensity theory, probabilities are tendencies or dispositions inherent in objects or situations.

[4]For instance, for $\{m_1, m_2, m_3\}$, $P = (1/3, 1/3, 1/3)$, $A = \{m_1, m_2\}$ and $B = \{m_3\}$ branching means that

$$S_{S,\,d}(1/3, 1/3, 1/3) = S_{S,\,d}(2/3, 1/3) + 2/3\, S_{S,\,d}(1/2, 1/2) + 1/3\, S_{S,\,d}(1). \tag{7}$$

[5]There are other axioms that uniquely characterise the Shannon entropy (cf. Klir 2006, section 3.2.2).

[6]We set $x \log[x] := 0$ for $x = 0$.

[7]For a discussion of the different interpretations of probability see, for instance, Howson (1995), Gillies (2000) and Mellor (2005).

The emphasis in information theory is on the receiver's amount of uncertainty about the next incoming message. This suggests that the $p(m_i)$ should be interpreted as epistemic probabilities (credences). While correct as a first stab, a more nuanced picture emerges once we ask the question of how the values of the $p(m_i)$ are set. Depending on how we understand the nature of the source, we obtain two very different answers. If the source itself is not probabilistic, then the $p(m_i)$ express the beliefs – and nothing but the beliefs – of receivers. For proponents of subjective probabilities these probabilities express the individual beliefs of an agent, and beliefs may vary between different receivers. Objectivist insists that all rational agents must come to the same value assignment. This can be achieved, for instance, by requiring that $S_{S,d}(P)$ be maximal, which singles out a unique distribution. This method, now known as Jaynes' maximum entropy principle, plays a role in statistical mechanics and will be discussed later.

Alternatively, the source source itself can be probabilistic. The probabilities associated with the source have to be ontic probabilities of one kind or other (frequencies, propensities, etc.). In this case agents are advised to use the so-called Principal Principle – roughly the rule that a rational agent's credence for a certain event to occur should be set equal to the objective probability (chance) of that event to occur.[8] In Shannon's setting this means that the $p(m_i)$ have to be equal to the source's objective probability of producing the message m_i. If this connection is established, the information transmitted in a channel is a measure of an objective property of a source.

It is worth emphasising that $S_{S,d}(P)$ is a technical conception of information which should not be taken as an analysis of the various senses 'information' has in ordinary discourse. In ordinary discourse information is often equated with knowledge, propositional content, or meaning. Hence 'information' is a property of a single message. Information as understood in information theory is not concerned with individual messages and their content; its focus is on *all* messages a source could possibly send. What makes a single message informative is not its meaning but the fact that it has been selected from a set of possible messages.

Given the probability distributions $P_m = (p_{m_1}, \ldots, p_{m_n})$ on $\{m_1, \ldots, m_n\}$, $P_s = (p_{s_1}, \ldots, p_{s_l})$ on $\{s_1, \ldots, s_l\}$, and the joint probability distribution $(p_{m_1,s_1}, p_{m_1,s_2} \ldots, p_{m_n,s_l})$[9] on $\{m_1s_1, m_1s_2, \ldots, m_ns_l\}$, the *conditional Shan-*

[8]The Principal Principle has been introduced by Lewis (1980); for a recent discussion see Frigg and Hoefer (2010).

[9]The outcomes m_i and s_j are not assumed to be independent.

non entropy is defined as

$$S_{S,\,d}(P_m \mid P_s) := \sum_{j=1}^{l} p_{s_j} \sum_{k=1}^{n} \frac{p_{m_k,s_j}}{p_{s_j}} \log[\frac{p_{m_k,s_j}}{p_{s_j}}]. \tag{9}$$

It measures the average information received from a message m_k given that a message s_j has been received before.

The Shannon entropy can be generalised to the continuous case. Let $p(x)$ be a probability density. The *continuous Shannon entropy* is

$$S_{S,\,c}(p) \;=\; -\int_{\mathbb{R}} p(x) \log[p(x)] dx \tag{10}$$

if the integral exists. The generalisation of (10) to densities of n variables $x_1, ..., x_n$ is straightforward. If $p(x)$ is positive, except for a set of Lebesgue measure zero, exactly on the interval $[a,\, b]$, $a, b \in \mathbb{R}$, then $S_{S,\,c}$ reaches its maximum, $\log[b-a]$, for $p(x) = 1/(b-a)$ in $[a,\, b]$ and zero elsewhere. Intuitively every change towards equalisation of $p(x)$ leads to an increase in entropy. For probability densities which are, except for a set of measure zero, positive everywhere on $\mathbb{R}$, the question of the maximum is more involved. If the standard deviation is held fixed at value σ, $S_{S,\,c}$ reaches its maximum for a Gaussian $p(x) = (1/\sqrt{2\pi}\sigma)\exp(-x^2/2\sigma^2)$, and the maximum value of the entropy is $\log[\sqrt{2\pi e}\sigma]$ (Ihara 1993, section 3.1; Shannon & Weaver 1949, pp. 88–89).

There is an important difference between the discrete and continuous Shannon entropy. In the discrete case, the value of the Shannon entropy is uniquely determined by the probability measure over the messages. In the continuous case the value depends on the coordinates we choose to describe the messages. Hence the continuous Shannon entropy cannot be regarded as measuring information, since an information measure should not depend on the way in which we describe a situation. But usually we are interested in entropy differences rather than in absolute values, and it turns out that *entropy differences* are coordinate independent and the continuous Shannon entropy can be used to measure differences in information (Ihara 1993, pp. 18–20; Shannon & Weaver 1949, pp. 90–91).[10]

We now turn to two other notions of information-theoretic entropy, namely Hartley's entropy and Rényi's entropy. The former preceded Shannon's entropy; the latter is a generalization of Shannon's entropy. One of

[10]This coordinate dependence reflects a deeper problem: the uncertainty reduced by receiving a message of a continuous distribution is infinite and hence not measured by $S_{S,\,c}$. Yet by approximating a continuous distribution by discrete distributions, one obtains that $S_{S,\,c}$ measures differences in information (Ihara 1993, p. 17).

the first account of information was introduced by Hartley (1928). Assume that $m := \{m_1, \ldots, m_n\}$, $n \in \mathbb{N}$, represents mutually exclusive possible alternatives and that one of the alternatives is true but we do not know which one. How can we measure the amount of information gained when knowing which of these n alternatives is true, or, equivalently, the uncertainty associated with these n possibilities? Hartley postulated that any function $S_H : \mathbb{N} \to \mathbb{R}^+$ answering this question has to satisfy the following axioms:

1. *Monotonicity.* The uncertainty increases with n: $S_H(n) \leq S_H(n+1)$ for all $n \in \mathbb{N}$.

2. *Branching.* The measure of information is independent of how the process is divided into parts: $S_H(n.m) = S_H(n)S_H(m)$, where '$n.m$' means that there are n times m alternatives.

3. *Normalization.* Per convention, $S_H(2) = 1$.

Again, there is exactly one function satisfying these axioms, namely $S_H(n) = \log[n]$ (Klir 2006, p. 26), which is now referred to as the *Hartley entropy*.

On the face of it this entropy is based solely on the concept of mutually exclusive alternatives, and it does not invoke probabilistic assumptions. However, views diverge on whether this is the full story. Those who deny this argue that the Hartley entropy implicitly assumes that all alternatives have *equal* weight. This amounts to assuming that they have equal probability, and hence the Hartley entropy is the special case of the Shannon entropy, namely the Shannon entropy for the uniform distribution. Those who deny this argue that Hartley's notion of alternatives does not presuppose probabilistic concepts and is therefore independent of Shannon's (cf. Klir 2006, pp. 25–30).

The Rényi entropies generalise the Shannon entropy. As with the Shannon entropy, assume a probability distribution $P = (p_1, ..., p_n)$ over $m = \{m_1, ..., m_n\}$. Require of a measure of information that it satisfies all the axioms of the Shannon entropy except for branching. Unlike the other axioms, it is unclear whether a measure of information should satisfy branching and hence whether it should be on the list of axioms (Rényi 1961). If the outcomes of two independent events with respective probabilities p and q are observed, we want the total received information to be the sum of the two partial informations. This implies that the amount of information received for message m_i is $-\log[p_i]$ (Jizba & Arimitsu 2004). If a weighted arithmetic mean is taken over the $-\log[p_i]$, we obtain the Shannon entropy. Now, is it possible to take another mean such that the remaining axioms about information are satisfied? If yes, these quantities are also possible measures of the

average information received. The general definition of a mean over $-\log[p_i]$ weighted by p_i is that it is of the form $f^{-1}(\sum_{i=1}^{n} p_i f(-\log[p_i]))$ where f is a continuous, strictly monotonic and invertible function. For $f(x) = x$ we obtain the Shannon entropy. There is only one alternative mean satisfying the axioms, namely $f(x) = 2^{(1-q)x}$, $q \in (0, \infty)$, $q \neq 1$. It corresponds to the *Rényi entropy of order q*:

$$S_{R,\,q}(P) := \frac{1}{1-q} \log[\sum_{k=1}^{n} p_k^q]. \tag{11}$$

The limit of the Rényi entropy for $q \to 1$ gives the Shannon entropy, i.e., $\lim_{q\to 1} S_{R,\,q}(P) = \sum_{k=1}^{n} -p_k \log[p_k]$ (Jizba & Arimitsu 2004; Rényi 1961), and for this reason one sets $S_{R,\,1}(P) := \sum_{k=1}^{n} -p_k \log[p_k]$.

4 Statistical Mechanics

Statistical mechanics (SM) aims to explain the behaviour of macroscopic systems in terms of the dynamical laws governing their microscopic constituents.[11] One of the central concerns of SM is to provide a micro-dynamical explanation of the Second Law of TD. The strategy to achieve this goal is to first introduce a mechanical notion of entropy, then to argue that it is in some sense equivalent to the TD entropy, and finally to show that it tends to increase if its initial value is low. There are two non-equivalent frameworks in SM, one associated with Boltzmann and one with Gibbs. In this section we discuss the various notions of entropy introduced within these frameworks and briefly indicate how they have been used to justify the Second Law.

SM deals with systems consisting of a large number of micro constituents. A typical example for such a system is a gas, which is made up of a large number n of particles of mass m confined to a vessel of volume V. And in this chapter we restrict attention to gases. Furthermore we assume that the system is isolated from its environment and hence that its total energy E is conserved. The behaviour of such systems is usually modeled by continuous measure-preserving dynamical systems. We discuss such systems in detail in the next section; for the time being it suffices to say that the phase space of the system is $6n$-dimensional, having three position and three momentum coordinates for every particle. This space is called the system's *γ-space* X_γ. x_γ denotes a vector in X_γ, and the x_γ are called *microstates*. X_γ is a direct product of n copies of the 6-dimensional phase space of one particle, called

[11] For an extended discussion of SM, see Frigg (2008), Sklar (1993) and Uffink (2006).

the particle's μ-space X_μ.[12] In what follows $x_\mu = (x, y, z, p_x, p_y, p_z)$ denotes a vector in X_μ; moreover, we use $\vec{r} = (x, y, z)$ and $\vec{p} = (p_x, p_y, p_z)$.[13]

In a seminal paper published in 1872 Boltzmann set out to show that the Second Law of TD is a consequence of the collisions between the particles of a gas. The distribution $f(x_\mu, t)$ specifies the fraction of particles in the gas whose position and momentum lies in the infinitesimal interval $(x_\mu,\ x_\mu + dx_\mu)$ at time t. In 1860 Maxwell showed that for a gas of *identical* and *non-interacting* particles in equilibrium the distribution had to be what is now called the *Maxwell-Boltzmann distribution*:

$$f(x_\mu, t) \ = \ \frac{\chi_V(\vec{r})\,(2\pi mkT)^{-3/2}}{\|V\|}\ \exp\left(-\frac{\vec{p}^{\,2}}{2mkT}\right), \tag{12}$$

where k is Boltzmann's constant, T the temperature of the gas, $\|V\|$ is the volume of the vessel, and $\chi_V(\vec{r})$ the characteristic function of volume V (it is 1 if $\vec{r} \in V$ and 0 otherwise).

The state of a gas at time t is fully specified by a distribution $f(x_\mu, t)$, and the dynamics of the gas can be studied by considering how this distribution evolves over time. To this end Boltzmann introduced the quantity

$$H_B(f) \ := \ \int_{X_\mu} f(x_\mu, t)\ \log[f(x_\mu, t)]\, dx_\mu \tag{13}$$

and set out to prove on the basis of mechanical assumptions about the collisions of gas molecules that $H_B(f)$ must decrease monotonically over the course of time and that it reaches its minimum at equilibrium where $f(x_\mu, t)$ becomes the Maxwell-Boltzmann distribution. This result, which is derived using the *Boltzmann equation*, is known as the *H-theorem* and is generally regarded as problematic.[14]

The problems of the H-theorem are not our concern. What matters is that the *fine-grained Boltzmann entropy* $S_{B,\,f}$ (also *continuous Boltzmann entropy*) is proportional to $H_B(f)$:

$$S_{B,\,f}(f) := -\,k\,n\,H_B(f). \tag{14}$$

Therefore, if the H-theorem were true, it would establish that the Boltzmann entropy increased monotonically and reached a maximum once the system's

[12]This terminology has been introduced by Ehrenfest & Ehrenfest (1912) and has been used since. The subscript 'μ' here stands for 'molecule' and has nothing to do with a measure.

[13]We use momentum rather than velocity since this facilitates the discussion of the connection of Boltzmann entropies with other entropies. One could also use the velocity $\vec{v} = \vec{p}/m$.

[14]See Emch & Liu (2002, pp. 92–105) and Uffink (2006, pp. 962–974).

distribution becomes the Maxwell-Boltzmann distribution. Thus, if we associated the Boltzmann entropy with the thermodynamic entropy, this would amount to a justification of the Second Law.

How are we to interpret the distribution $f(x_\mu, t)$? As introduced, $f(x_\mu, t)$ reflects the distribution of the particles: it tells what fraction of the particles in the gas are located in a certain region of the phase space. So it can be interpreted as an (approximate) actual distribution, involving no probabilistic notions. But $f(x_\mu, t)$ can also be interpreted probabilistically, as specifying the probability that a particle drawn at random from the gas is located in a particular part of the phase space. This probability is most naturally interpreted in a frequentist way: if we keep drawing molecules at random from the gas, then $f(x_\mu, t)$ gives us the relative frequency of molecules drawn from a certain region of phase space.

In 1877 Boltzmann presented an altogether different approach to justifying the Second Law.[15] Since energy is conserved and the system is confined to volume V, each state of a particle lies within a finite sub-region $X_{\mu,a}$ of X_μ, the accessible region of X_μ. Now we *coarse-grain* $X_{\mu,a}$, i.e., we choose a partition $\omega = \{\omega_i : i = 1, \ldots, l\}$ of $X_{\mu,a}$.[16] The cells ω_i are taken to be rectangular with respect to the position and momentum coordinates and of equal volume $\delta\omega$, i.e., $\mu(\omega_i) = \delta\omega$, for all $i = 1, \ldots, l$, where μ is the Lebesgue measure on the 6-dimensional phase space of one particle. The *coarse-grained microstate*, also called *arrangement*, is a specification of which particle's state lies in which cell of ω.

The macroscopic properties of a gas (e.g., temperature, pressure) do not depend on which specific molecule is in which cell of the partition but are determined solely by the number of particles in each cell. A specification of how many particles are in each cell is called a *distribution* $D = (n_1, \ldots, n_l)$, meaning that n_1 particles are in cell ω_1, etc. Clearly, $\sum_{j=1}^{l} n_j = n$. We label the different distributions with a discrete index i and denote the i^{th} distribution by D_i. D_i/n can be interpreted in the same way as $f(x_\mu, t)$ above.

Several arrangements correspond to the same distribution. More precisely, elementary combinatorial considerations show that

$$G(D) := \frac{n!}{n_1! \ldots n_l!} \tag{15}$$

arrangements are compatible with a given distribution D. The so-called

[15] See Uffink (2006, 974–983) and Frigg (2008, 107–113). Frigg (2009*a*, 2009*b*) provides a discussion of Boltzmann's use of probabilities.

[16] We give a rigorous definition of a partition in the next section.

coarse-grained Boltzmann entropy (also *combinatorial entropy*) is defined as:

$$S_{B,\omega}(D) := k \log[G(D)]. \tag{16}$$

Since $G(D)$ is the number of arrangements compatible with a given distribution and the logarithm is a monotonic function, $S_{B,\omega}(D)$ is a measure for the number of arrangements that are compatible with a given distribution: the greater $S_{B,\omega}(D)$, the more arrangements are compatible with a given distribution. Hence $S_{B,\omega}(D)$ is a measure of how much we can infer about the arrangement of a system on the basis of its distribution. The higher $S_{B,\omega}(D)$, the less information a distribution confers about the arrangement of the system.

Boltzmann then postulated that the distribution with the highest entropy was the equilibrium distribution, and that systems had a natural tendency to evolve from states of low to states of high entropy. However, as later commentators, most notably Ehrenfest & Ehrenfest (1912), pointed out, for the latter to happen further dynamical assumptions (e.g., ergodicity) are needed. If such assumptions are in place, the n_i evolve so that $S_{B,\omega}(D)$ increases and then stays close to its maximum value most of the time (Lavis 2004, 2008).

There is a third notion of entropy in the Boltzmannian framework, and this notion is preferred by contemporary Boltzmannians.[17] We now consider X_γ rather than X_μ. Since there are constraints on the system, its state will lie within a finite sub-region $X_{\gamma,a}$ of X_γ, the accessible region of X_γ.[18]

If the gas is regarded as a macroscopic object rather than as a collection of molecules, its state can be characterised by a small number of macroscopic variables such as temperature, pressure and density. These values are then usually coarse-grained so that all values falling into a certain range are regarded as belonging to the same macrostate. Hence the system can be described as being in one of a finite number of macrostates M_i, $i = 1, \ldots, m$. The set of M_i is complete in that at any given time t the system must be in exactly one M_i. It is a basic posit of the Boltzmann approach that a system's macrostate supervenes on its fine-grained microstate, meaning that a change in the macrostate must be accompanied by a change in the fine-grained microstate. Therefore, to every given microstate x_γ there corresponds *exactly*

[17] See, for instance, Goldstein (2001) and Lebowitz (1999).

[18] These constrains include conservation of energy. Therefore, $\Gamma_{\gamma,a}$ is $(6n-1)$-dimensional. This causes complications because the measure μ needs to be restricted to the $(6n-1)$-dimensional energy hypersurface and the definitions of macroregions become more complicated. In order to keep things simple, we assume that $\Gamma_{\gamma,a}$ is $6n$-dimensional. For the (6n-1)-dimensional case, see Frigg (2008, pp. 107–114).

one macrostate $M(x_\gamma)$. But many different microstates can correspond to the same macrostate. We therefore define

$$X_{M_i} := \{x_\gamma \in X_{\gamma,a} \,|\, M_i = M(x_\gamma)\} \ , \ i = 1, ..., m, \tag{17}$$

which is the subset of $X_{\gamma,a}$ consisting of all microstates that correspond to macrostate M_i. The X_{M_i} are called *macroregions*. Clearly, they form a partition of $X_{\gamma,a}$.

The Boltzmann entropy of a macrostate M is[19]

$$S_{B,m}(M) := k \log[\mu(X_M)]. \tag{18}$$

Hence $S_{B,m}(M)$ measures the portion of the system's γ-space that is taken up by microstates that correspond to M. Consequently, $S_{B,m}(M)$ measures how much we can infer about where in γ-space the system's microstate lies: the higher $S_{B,m}(M)$, the larger the portion of the γ-space in which the system's microstate could be.

Given this notion of entropy, the leading idea is to argue that the dynamics of systems is such that $S_{B,m}$ increases. Most contemporary Boltzmannians aim to achieve this by arguing that entropy increasing behaviour is *typical*; see, for instance, Goldstein (2001). These arguments are the subject of ongoing controversy (see Frigg 2009*a*, 2009*b*).

We now turn to a discussion of the interrelationships between the various entropy notions introduced so far. Let us begin with $S_{B,\omega}$ and $S_{B,m}$. $S_{B,\omega}$ is a function of a distribution over a partition of $X_{\mu,a}$, while $S_{B,m}$ takes cells of a partition of $X_{\gamma,a}$ as arguments. The crucial point to realise is that each distribution corresponds to a well-defined region of $X_{\gamma,a}$: the choice of a partition of $X_{\mu,a}$ induces a partition of $X_{\gamma,a}$ (because X_γ is the Cartesian product of n copies of X_μ). Hence any D_i uniquely determines a region X_{D_i} so that all states $x_\gamma \in X_{D_i}$ have distribution D_i:

$$X_{D_i} := \{x_\gamma \in X_\gamma \,|\, D(x_\gamma) = D_i\}, \tag{19}$$

where $D(x_\gamma)$ is the distribution of state x_γ. Because all cells have measure $\delta\omega$, equations (15) and (19) imply:

$$\mu(X_{D_i}) = G(D_i)\,(\delta\omega)^n. \tag{20}$$

Given this, the question of the relation between $S_{B,\omega}$ and $S_{B,m}$ comes down to the question of how the X_{D_i} and the X_{M_i} relate. Since there are no standard procedures to construct the X_{M_i}, one can use the above considerations about how distributions determine regions to construct the X_{M_i},

[19] See, e.g., Goldstein (2001, p. 43) and Lebowitz (1999, p. 348).

making $X_{D_i} = X_{M_i}$ true by definition. So one can say that $S_{B,\omega}$ is a special case of $S_{B,m}$ (or that it is a concrete realisation of the more abstract notion of $S_{B,m}$). If $X_{D_i} = X_{M_i}$, equations (18) and (20) imply:

$$S_{B,m}(M_i) \;=\; k \log[G(D_i)] + k\, n \log[\delta\omega]. \tag{21}$$

Hence $S_{B,m}(M_i)$ equals $S_{B,\omega}$ up to an additive constant.

How do $S_{B,m}$ and $S_{B,f}$ relate? Assume that $X_{D_j} = X_{M_j}$, that the system is large, and that there are many particles in each cell ($n_j \gg 1$ for all j), which allows us to use Stirling's formula: $n! \approx \sqrt{2\pi n}(n/e)^n$. Plugging equation (15) into equation (21), yields (Tolman 1938, chapter 4)

$$\log[\mu(X_{M_j})] \approx n \log[n] - \sum_{i=1}^{l} n_i \log[n_i] + n \log[\delta\omega]. \tag{22}$$

Clearly, for the n_i used in the definition of $S_{B,\omega}$ we have

$$n_i \approx \tilde{n}_i(t) := n \int_{\omega_i} f(x_\mu, t)\, dx_\mu. \tag{23}$$

Unlike the n_i the $\tilde{n}_i$ need not be integers. If $f(x_\mu, t)$ does not vary much in each cell ω_i, we find:

$$\sum_{i=1}^{l} n_i \log[n_i] \approx n H_B + n \log[n] + n \log[\delta\omega]. \tag{24}$$

Comparing (22) and (24) yields $-nkH_B \approx k \log[\mu(X_{M_j})]$, i.e., $S_{B,m} \approx S_{B,f}$. Hence, for a large number of particles $S_{B,m}$ and $S_{B,f}$ are approximately equal.

How do $S_{B,m}$ and the Shannon entropy relate? According to equation (22),

$$S_{B,m}(M_j) \;\approx -\, k \sum_{i=1}^{l} n_i \log[n_i] + C(n,\, \delta\omega), \tag{25}$$

where $C(n,\, \delta\omega)$ is a constant depending on n and $\delta\omega$. Introducing the quotients $p_j := n_j/n$, we find

$$S_{B,m}(M_j) \;\approx -n\, k \sum_{i=1}^{l} p_i \log[p_i] + \tilde{C}(n,\, \delta\omega), \tag{26}$$

where $\tilde{C}(n,\, \delta\omega)$ is a constant depending on n and $\delta\omega$. The quotients p_i are finite relative frequencies for a particle being in ω_i. The p_i can be interpreted

as the probability of finding a randomly chosen particle in cell ω_i. Then, if we regard the ω_i as messages, $S_{B,\,m}(M_i)$ is equivalent to the Shannon entropy up to the multiplicative constant nk and the additive constant $\tilde{C}$.

Finally, how does $S_{B,\,f}$ relate to the TD entropy? The TD entropy of an ideal gas is given by the so-called Sackur-Tetrode formula

$$S_{TD} \;=\; n\,k \log\left[\left(\frac{T}{T_0}\right)^{3/2}\left(\frac{V}{V_0}\right)\right], \tag{27}$$

where T_0 and V_0 are the temperature and the volume of the gas at reference point E (see Reiss 1965, pp. 89–90). One can show that $S_{B,\,f}$ for the Maxwell-Boltzmann distribution is equal to equation (27) up to an additive constant (Emch & Liu 2002, p. 98; Uffink 2006, p. 967). This is an important result. However, it is an open question whether this equivalence holds for systems with interacting particles, that is, for systems different from ideal gases.

The object of study in the Gibbs approach is not an individual system (as in the Boltzmann approach) but an ensemble – an uncountably infinite collection of independent systems that are all governed by the same equations but whose states at a time t differ. The ensemble is specified by an everywhere positive density function $\rho(x_\gamma, t)$ on the system's γ-space: $\rho(x_\gamma, t)dx_\gamma$ is the infinitesimal fraction of systems in the ensemble whose state lies in the $6n$-dimensional interval $(x_\gamma, x_\gamma + dx_\gamma)$. The time evolution of the ensemble is then associated with changes in the density function in time.

$\rho(x_\gamma, t)$ is a probability density, reflecting the probability at time t of finding the state of a system in region $R \subseteq X_\gamma$:

$$p_t(R) \;=\; \int_R \rho(x_\gamma, t)dx_\gamma. \tag{28}$$

The *fine-grained Gibbs* entropy (also *ensemble entropy*) is defined as:

$$S_{G,\,f}(\rho) \;:=\; -k\int_{X_\gamma} \rho(x_\gamma, t)\log[\rho(x_\gamma, t)]dx_\gamma. \tag{29}$$

How to interpret $\rho(x_\gamma, t)$ (and hence $p_t(R)$) is far from clear. Edwin Jaynes proposed to interpret $\rho(x_\gamma, t)$ epistemically; we turn to his approach to SM below. There are (at least) two possible ontic interpretations: a frequency interpretation and a time average interpretation. On the frequency interpretation one thinks about an ensemble as analogous to an urn, but rather than containing balls of different colours the ensemble contains systems in different micro-states (Gibbs 1981, p. 163). The density $\rho(x_\gamma, t)$ specifies the frequency with which we draw systems in a certain micro-state. On

the time average interpretation, $\rho(x_\gamma, t)$ reflects the fraction of time that the system would spend, in the long run, in a certain region of the phase space if it was left to its own. Although plausible at first blush, both interpretations face serious difficulties and it is unclear whether these can be met (see Frigg 2008, pp. 153–155).

If we regard $S_{G,f}(\rho)$ as equivalent to the TD entropy (which is common), then $S_{G,f}(\rho)$ is expected to increase over time (during an irreversible adiathermal process) and assumes a maximum in equilibrium. However, systems in SM are Hamiltonian, and it is a consequence of an important theorem of Hamiltonian mechanics, *Liouville's theorem*, that $S_{G,f}$ is a constant of motion: $dS_{G,f}/dt = 0$. So $S_{G,f}$ remains constant, and hence the approach to equilibrium cannot be described in terms of an increase in $S_{G,f}$.

The standard way to solve this problem is to consider the coarse-grained Gibbs entropy instead. This solution has been suggested by Gibbs (1981, chapter 12) and has since been endorsed by many (e.g., Penrose 1970). Consider a partition ω of X_γ where the cells ω_i are of equal volume $\delta\omega$. The *coarse-grained density* $\bar{\rho}(x_\gamma, t)$ is defined as the density that is uniform within each cell, taking as its value the average value in this cell:

$$\bar{\rho}_\omega(x_\gamma, t) := \frac{1}{\delta\omega} \int_{\omega(x_\gamma)} \rho(x'_\gamma, t) dx'_\gamma, \tag{30}$$

where $\omega(x_\gamma)$ is the cell in which x_γ lies. We can now define the *coarse-grained Gibbs entropy*:

$$S_{G,\omega}(\rho) := S_{G,f}(\bar{\rho}_\omega) = -k \int_{X_\gamma} \bar{\rho}_\omega \log[\bar{\rho}_\omega] dx_\gamma. \tag{31}$$

One can prove that $S_{G,\omega} \geq S_{G,f}$; the equality holds iff the fine grained distribution is uniform over the cells of the coarse-graining (see Lavis 2004, p. 229; Wehrl 1978, p. 672). The coarse-grained density $\bar{\rho}_\omega$ is not subject to Liouville's theorem and is not a constant of motion. So $\bar{\rho}_\omega$ could, in principle, increase over time.[20]

How do the two Gibbs entropies relate to the other notions of entropy introduced so far? The most straightforward connection is between the Gibbs entropy and the continuous Shannon entropy, which differ only by the multiplicative constant k. This realisation provides a starting point for Jaynes's (1983) information-based interpretation of SM, at the heart of which lies a radical reconceptualisation of SM. On his view, SM is about our knowledge of the world, not about the world. The probability distribution represents

[20]There is a thorny issue under which conditions the coarse-grained entropy actually increases (see Lavis 2004).

our state of knowledge about the system and not some matter of fact about the system: $\rho(x_\gamma, t)$ represents our lack of knowledge about a micro-state of a system given its macro condition and entropy is a measure of how much knowledge we lack.

Jaynes then postulated that to make predictions we should always use the distribution that maximises uncertainty under the given macroscopic constraints. This means that we are asked to find the distribution for which the the Gibbs entropy is maximal, and then use this distribution to calculate expectation values of the variables of interest. This prescription is now know as *Jaynes' Maximum Entropy Principle*. Jaynes could show that this principle recovers the standard SM distributions (e.g., the microcanonical distribution for isolated systems).

The idea behind this principle is that we should always choose the distribution that is maximally non-committal with respect to the missing information because by not doing so we would make assertions for which we have no evidence. Although intuitive at first blush, the maximum entropy principle is fraught with controversy (see, for instance, Howson and Urbach 2006, pp. 276–288).[21]

A relation between $S_{G,f}(\rho)$ and the TD entropy can be established only case by case. $S_{G,f}(\rho)$ coincides with S_{TD} in relevant cases arising in practice. For instance, the calculation of the entropy of an ideal gas from the microcanonical ensemble yields equation (27) – up to an additive constant (Kittel 1958, p. 39).

Finally, how do the Gibbs and Boltzmann entropies relate? Let us start with the fine grained entropies $S_{B,f}$ and $S_{G,f}$. Assume that the particles are identical and non-interacting. Then $\rho(x_\gamma, t) = \prod_{i=1}^{n} \rho_i(x_\mu, t)$, where ρ_i is the density pertaining to particle i. Then

$$S_{G,f}(\rho) := -k\, n \int_{X_\mu} \rho_1(x_\mu, t) \log[\rho_1(x_\mu, t)] dx_\mu, \tag{32}$$

which is *formally* equivalent to $S_{B,f}$ (14). The question is how ρ_1 and f relate since they are different distributions. f is the distribution of n particles over the phase space; ρ_1 is a one particle function. Because the particles are identical and noninteracting, we can apply the law of large numbers to conclude that it is very likely that the probability of finding a given particle

[21] For a discussion of Jaynes's take on non-equilibrium SM, see Sklar (1993, pp. 255–257). Furthermore, Tsallis (1988) proposed a way of deriving the main distributions of SM which is very similar to Jaynes' based on establishing a connection between what is now called the Tsallis entropy and the Rényi entropy. A similar attempt using only the Rényi entropy has been undertaken by Bashkirov (2006).

in a particular region of phase space is approximately equal to the proportion of particles in that region. Hence $\rho_1 \approx f$ and $S_{G,f} \approx S_{B,f}$.

A similar connection exists between the coarse grained entropies $S_{G,m}$ and $S_{B,\omega}$. If the particles are identical and non-interacting, one finds

$$S_{G,\omega} = -k\,n \sum_{i=1}^{l} \int_{\omega_i} \frac{\Omega_i}{\delta\omega} \log[\frac{\Omega_i}{\delta\omega}] dx_\mu = -k\,n \sum_{i=1}^{l} \Omega_i \log[\Omega_i] + C(n, \delta\omega), \quad (33)$$

where $\Omega_i = \int_{\omega_i} \rho_1 dx_\mu$. This is *formally* equivalent to $S_{B,m}$ (26), which in turn is equivalent (up to an additive constant) to $S_{B,\omega}$ (16). Again for large n we can apply the law of large numbers to conclude that it is very likely that $\Omega_i \approx p_i$ and $S_{G,m} = S_{B,\omega}$.

It is crucial for the connections between the Gibbs and the Boltzmann entropy that the particles are identical and noninteracting. It is unclear whether the conclusions hold if these assumptions are relaxed.[22]

5 Dynamical Systems Theory

In this section we focus on the main notions of entropy in dynamical systems theory, namely the Kolmogorov-Sinai entropy (KS-entropy) and the topological entropy.[23] They occupy centre stage in *chaos theory* – a mathematical theory of deterministic yet irregular and unpredictable or even random behaviour.[24]

We begin by briefly recapitulating the main tenets of dynamical systems theory.[25] The two main elements of every dynamical system are a set X of all possible states x, the *phase space* of the system, and a family of transformations $T_t : X \to X$ mapping the phase space to itself. The parameter t is time, and the transformations $T_t(x)$ describe the *time evolution* of the system's instantaneous state $x \in X$. For the systems we have discussed in the last section X consists of the positions and momenta of all particles in the system and T_t is the time evolution of the system under the dynamical laws. If t is a positive real number or zero (i.e., $t \in \mathbb{R}_0^+$), the system's

[22]Jaynes (1965) argues that the Boltzmann entropy differs from the Gibbs entropy except for noninteracting and identical particles. However, he *defines* the Boltzmann entropy as (32). As argued, (32) is equivalent to the Boltzmann entropy if the particles are identical and noninteracting, but this does not appear to be generally the case. So Jaynes's (1965) result seems useless.

[23]There are also a few other less important entropies in dynamical systems theory, e.g., the Brin-Katok local entropy (see Mañé 1987).

[24]For a discussion of the kinds or randomness in chaotic systems, see Berkovitz, Frigg & Kronz (2006) and Werndl (2009*a*, 2009*b*, 2009*d*).

[25]For more details, see Cornfeld, Fomin & Sinai (1982) and Petersen (1983).

dynamics is *continuous*. If t is a natural number or zero (i.e., $t \in \mathbb{N}_0$), its dynamics is *discrete*.[26] The family T_t defining the dynamics must have the structure of a semi-group where $T_{t_1+t_2}(x) = T_{t_2}(T_{t_1}(x))$ for all t_1, t_2 either in $\mathbb{R}_0^+$ (continuous time) or $\mathbb{N}_0$ (discrete time).[27] The continuous trajectory through x is the set $\{T_t(x) \mid t \in \mathbb{R}_0^+\}$; the discrete trajectory through x is the set $\{T_t(x) \mid t \in \mathbb{N}_0\}$.

Continuous time evolutions often arise as solutions to differential equations of motion (such as Newton's or Hamilton's). In dynamical systems theory the class of allowable equations of motion is usually restricted to ones for which solutions exist and are unique for all times $t \in \mathbb{R}$. Then $\{T_t : t \in \mathbb{R}\}$ is a group where $T_{t_1+t_2}(x) = T_{t_2}(T_{t_1}(x))$ for all $t_1, t_2 \in \mathbb{R}$ and are often called *flows*. In what follows we only consider continuous systems that are flows.

For discrete systems the maps defining the time evolution neither have to be injective nor surjective, and so $\{T_t : t \in \mathbb{N}_0\}$ is only a semigroup. All T_t are generated as iterative applications of the single map $T_1 := T : X \to X$ because $T_t = T^t$, and we refer to the $T_t(x)$ as *iterates of* x. Iff T is invertible, T_t is defined both for positive and negative times and $\{T_t : t \in \mathbb{Z}\}$ is a group.

It follows that all dynamical systems are *forward-deterministic*: any two trajectories that agree at one instant of time agree at all *future* times. If the dynamics of the system is invertible, the system is deterministic *tout court*: any two trajectories that agree at one instant of time agree at all times (Earman 1971).

Two kinds of dynamical systems are relevant for our discussion, measure-theoretical and topological dynamical ones. A *topological dynamical system* has a metric defined on X.[28] More specifically, a *discrete topological dynamical system* is a triple (X, d, T) where d is a metric on X and $T : X \to X$ is a mapping. *Continuous topological dynamical systems* (X, d, T_t), $t \in \mathbb{R}$, are defined accordingly where T_t is the above semi-group. Topological systems allow for a wide class of dynamical laws since the T_t have to be neither injective nor surjective.

A *measure-theoretical dynamical system* is one whose phase space is endowed with a measure.[29] More specifically, a *discrete measure-theoretical dynamical system* (X, Σ, μ, T) consists of a phase space X, a σ-algebra Σ

[26]The reason not to choose $t \in \mathbb{Z}$ is that some maps, e.g., the logistic map, are not invertible.

[27]$S = \{a, b, c, ...\}$ is a *semigroup* iff there is a multiplication operation '$\cdot$' on S so that (i) $a \cdot b \in S$ for all $a, b \in S$; (ii) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b, c \in S$; (iii) $e \cdot a = a \cdot e = a$ for all $a \in S$. A semigroup as defined here is also called a *monoid*. If for every $a \in S$ there is a $a^{-1} \in S$ so that $a^{-1} \cdot a = a \cdot a^{-1} = e$, S is a *group*.

[28]For a discussion of metrics, see Sutherland (2002).

[29]See Halmos (1950) for an introduction to measures.

on X, a measure μ, and a measurable transformation $T : X \to X$. If T is *measure-preserving*, i.e., $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \Sigma$ where $T^{-1}(A) := \{x \in X : T(x) \in A\}$, we have a *discrete measure-preserving dynamical system*. It makes only sense to speak of measure-preservation if T is surjective. Therefore, we suppose that the T in measure-preserving systems is surjective. However, we do not presuppose that it is injective because important maps are not injective, e.g., the logistic map.

A *continuous measure-theoretical dynamical system* is a quadruple (X, Σ, μ, T_t), where $\{T_t : t \in \mathbb{R}_0^+\}$ is the above semigroup of transformations which are measurable on $X \times \mathbb{R}_0^+$, and the other elements are as above. Such a system is a *continuous measure-preserving dynamical system* if T_t is measure preserving for all t (again, we presuppose that all T_t are surjective).

We make the (common) assumption that the measure of measure-preserving systems is normalised: $\mu(X) = 1$. The motivation for this is that normalised measures are probabilitiy measures, making it possible to use probability calculus. This raises the question of how to interpret these probabilities. This issue is particularly thorny because it is widely held that there cannot be ontic probabilities in deterministic systems: either the dynamics of a system is deterministic or chancy, but not both. This dilemma can be avoided if one interprets probabilities epistemically, i.e., as reflecting lack of knowledge. This is what Jaynes did in SM. Although sensible in some situations, this interpretation is clearly unsatisfactory in others. Roulette wheels and dice are paradigmatic examples of chance setups, and it is widely held that there are ontic chances for certain events to occur: the chance of getting a '3' when throwing a dice is 1/6, and this is so due to how the world *is* and it has nothing to do with what we happen to *know* about it. Yet, from a mechanical point of view these are deterministic systems. Consequently, there must be ontic interpretations of probabilities in deterministic systems. There are at least three options available. The first is the time average interpretation already mentioned above: the probability of an event E is the fraction of time that the system spends (in the long run) in the region of X associated with E (Falconer 1990, p. 254; Werndl 2009*d*). The ensemble interpretation defines the measure of a set A at time t as the fraction of solutions starting from some set of initial conditions that are in A at t. A third option is the so-called Humean Best System analysis originally proposed by Lewis (1980). Roughly speaking, this interpretation is an elaboration of (finite) frequentism. Lewis' own assertions notwithstanding, this interpretation works in the context of deterministic systems (Frigg and Hoefer 2010).

Let us now discuss the notions of volume-preservation and measure-preservation. If the preserved measure is the Lebesgue measure, the system is *volume-preserving*. If the system fails to be volume-preserving, then it is

dissipative. Being dissipative is not the failure of measure preservation with respect to *any* measure (as a common misconception has it); it is preservation with respect to the *Lebesgue measure*. In fact many dissipative systems preserve measures. More precisely, if (X, Σ, λ, T) or $(X, \Sigma, \lambda, T_t)$ is dissipative (λ is the Lebesgue measure), often, although not always, there exists a measure $\mu \neq \lambda$ such that (X, Σ, μ, T) or (X, Σ, μ, T_t) is measure-preserving. The Lorenz system and the logistic maps are cases in point.

A *partition* $\alpha = \{\alpha_i \,|\, i = 1, \ldots, n\}$ of (X, Σ, μ) is a collection of non-empty, non-intersecting measurable sets that cover X: $\alpha_i \cap \alpha_j = \emptyset$ for all $i \neq j$ and $X = \bigcup_{i=1}^{n} \alpha_i$. The α_i are called *atoms*. If α is a partition, $T_t^{-1}\alpha := \{T_t^{-1}\alpha_i \,|\, i = 1, \ldots, n\}$ is a partition too. $T_t\alpha := \{T_t\alpha_i \,|\, i = 1, \ldots, n\}$ is a partition iff T_t is invertible. Given two partitions $\alpha = \{\alpha_i \,|\, i = 1, \ldots, n\}$ and $\beta = \{\beta_j \,|\, j = 1, \ldots, m\}$, the *join* $\alpha \vee \beta$ is defined as $\{\alpha_i \cap \beta_j \,|\, i = 1, \ldots, n; j = 1, \ldots, m\}$.

This concludes our brief recapitulation of dynamical systems theory. The rest of this section concentrates on measure preserving systems. This is not very restrictive because many systems, including all deterministic Newtonian systems, many dissipative systems and all chaotic systems (Werndl 2009*d*), fall into this class.

Let us first discuss the KS-entropy. Given a partition $\alpha = \{\alpha_1, \ldots, \alpha_k\}$, let $H(\alpha) := -\sum_{i=1}^{k} \mu(\alpha_i) \log[\mu(\alpha_i)]$. For a discrete system (X, Σ, μ, T) consider

$$H_n(\alpha, T) := \frac{1}{n} H(\alpha \vee T^{-1}\alpha \vee \ldots \vee T^{-n+1}\alpha). \tag{34}$$

The limit $H(\alpha, T) := \lim_{n\to\infty} H_n(\alpha, T)$ exists, and the KS-entropy is defined as (Petersen 1983, p. 240):

$$S_{KS}(X, \Sigma, \mu, T) := \sup_{\alpha}\{H(\alpha, T)\}. \tag{35}$$

For a continuous system (X, Σ, μ, T_t) it can be shown that for any t_0, $-\infty < t_0 < \infty$, $t_0 \neq 0$,

$$S_{KS}(X, \Sigma, \mu, T_{t_0}) = |t_0| S_{KS}(X, \Sigma, \mu, T_1), \tag{36}$$

where $S_{KS}(X, \Sigma, \mu, T_{t_0})$ is the KS-entropy of the discrete system $(X, \Sigma, \mu, T_{t_0})$ and $S_{KS}(X, \Sigma, \mu, T_1)$ is the KS-entropy of the discrete system (X, Σ, μ, T_1) (Cornfeld et al. 1982). Consequently, the KS-entropy of a continuous system (X, Σ, μ, T_t) is defined as $S_{KS}(X, \Sigma, \mu, T_1)$, and when discussing the meaning of the KS-entropy it suffices to focus on (35).[30]

[30]For experimental data the KS-entropy, and also the topological entropy (discussed later), is rather hard to determine. For details, see Eckmann & Ruelle (1985), and Ott (2002); see also Shaw (1985), who discusses how to define a quantity similar to the KS-entropy for dynamical systems with added noise.

How can the KS-entropy be interpreted? There is a fundamental connection between dynamical systems theory and information theory. For a dynamical system (X, Σ, μ, T) each $x \in X$ produces, relative to a partition α, an infinite string of messages $m_0 m_1 m_2 \ldots$ in an alphabet of k letters via the coding $m_j = \alpha_i$ iff $T^j(x) \in \alpha_i$, $j \geq 0$. Assume that (X, Σ, μ, T) is interpreted as the source. Then the output of the source are the strings $m_0 m_1 m_2 \ldots$. If the measure is interpreted as probability density, we have a probability distribution over these strings. Hence the whole apparatus of information theory can be applied to these strings.[31] In particular, notice that $H(\alpha)$ is the Shannon entropy of $P = (\mu(\alpha_1), \ldots, \mu(\alpha_k))$ and measures the average information of the message α_i.

In order to motivate the KS-entropy, consider for $\alpha := \{\alpha_1, \ldots, \alpha_k\}$ and $\beta := \{\beta_1, \ldots, \beta_l\}$:

$$H(\alpha \,|\, \beta) := \sum_{j=1}^{l} \mu(\beta_j) \sum_{i=1}^{k} \frac{\mu(\alpha_i \cap \beta_j)}{\mu(\beta_j)} \log[\frac{\mu(\alpha_i \cap \beta_j)}{\mu(\beta_j)}]. \tag{37}$$

Recalling the definition of the conditional Shannon entropy (9), we see that $H(\alpha \,|\, \vee_{k=1}^{n} T^{-k}\alpha)$ measures the average information received about the present state of the system whatever n past states have been already recorded. It is proven that (Petersen 1983, pp. 241–242):

$$S_{KS}(X, \Sigma, \mu, T) = \sup_{\alpha}\{\lim_{n\to\infty} H(\alpha \,|\, \vee_{k=1}^{n} T^{-k}\alpha)\}. \tag{38}$$

Hence the KS-entropy is linked to the Shannon entropy, namely it measures the highest average information received about the present state of the system relative to a coding α given the past states that have been received.

Clearly, equation (38) implies that

$$S_{KS}(X, \Sigma, \mu, T) = \sup_{\alpha}\{\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} H(\alpha| \vee_{i=1}^{k} T^{-i}\alpha)\}. \tag{39}$$

Hence the KS-entropy can be also interpreted, as Frigg (2004, 2006) does, as the highest average of the average information gained about the present state of the system relative to a coding α whatever past states have been received.

This is not the only connection to the Shannon entropy: let us regard *strings of length* n, $n \in \mathbb{N}$, produced by the dynamical system relative to a coding α as messages. The probability distribution of these possible strings of

[31] For details, see Frigg (2004) and Petersen (1983, pp. 227–234).

length n relative to α is $\mu(\beta_i)$, $1 \leq i \leq h$, $\beta = \{\beta_1, \ldots, \beta_h\} := (\alpha \vee T^{-1}\alpha \vee \ldots \vee T^{-n+1}\alpha)$. Hence $H_n(\alpha, T)$ measures the average amount of information which the system produces *per step* over the first n steps relative to the coding α, and $\lim_{n\to\infty} H_n(\alpha, T)$ measures the average amount of information produced by the system per step relative to α. Consequently, $\sup_\alpha\{H(\alpha, T)\}$ measures the highest average amount of information that the system can produce per step relative to a coding (cf. Petersen 1983, pp. 227–234).

A positive KS-entropy is often linked to chaos. The interpretations discussed provide a rational for this: the Shannon information measures uncertainty, and this uncertainty is a form of unpredictability (Frigg 2004). Hence a positive KS-entropy means that relative to some codings the behaviour of the system is unpredictable.

Kolmogorov (1958) was the first to connect dynamical systems theory with information theory. Based on Kolmogorov's work, Sinai (1959) introduced the KS-entropy. One of Kolmogorov's main motivations was the following.[32] Kolmogorov conjectured that while the deterministic systems used in science produce no information, the stochastic processes used in science produce information, and the KS-entropy was introduced to capture the property of producing positive information. It was a big surprise when it was found that also several deterministic systems used in science, e.g., some Newtonian systems etc., have positive KS-entropy. Hence this attempt of separating deterministic systems from stochastic processes failed (Werndl 2009*a*).

Due to lack of space we cannot discuss another, quite different, interpretation of the Kolmogorov-Sinai entropy, where $\sup_\alpha\{H(\alpha, T)\}$ is a measure of the highest average rate of exponential divergence of solutions relative to a partition as time goes to infinity (Berger 2001, pp. 117–118). This implies that if $S_{KS}(X, \Sigma, \mu, T) > 0$, there is exponential divergence and thus unstable behaviour on some regions of phase space, explaining the link to chaos. This interpretation does *not* require that the measure is interpreted as probability.

There is also another connection of the KS-entropy to exponential divergence of solutions. The Lyapunov exponents of x measure the mean exponential divergence of solutions originating near x, where the existence of positive Lyapunov exponents indicates that, in some directions, solutions diverge exponentially on average. *Pesins's theorem* states that under certain assumptions $S_{KS}(X, \Sigma, \mu, T) = \int_X S(x)d\mu$, where $S(x)$ is the sum of the positive Liapunov exponents of x. Another important theorem which should be mentioned is *Brudno's theorem*, which states that if the system is ergodic

[32]Another main motivation was to make progress on the problem of which systems are probabilistically equivalent (Werndl 2009*c*).

and certain other conditions hold, $S_{KS}(X, \Sigma, \mu, T)$ equals the algorithmic complexity (a measure of randomness) of almost all solutions (Batterman & White 1996).

The interpretations of the KS-entropy as measuring exponential divergence are not connected to any other notion of entropy or to what entropy notions are often believed to capture, such as information (Grad 1961, pp. 323–234; Wehrl 1978, pp. 221–224). To conclude, the only link of the KS-entropy to entropy notions is with the Shannon entropy.

Let us now discuss the topological entropy, which is always defined only for discrete systems. It was first introduced by Adler, Konheim & McAndrew (1965); later Bowen (1971) introduced two other equivalent definitions.

We first turn to Adler et al.'s definition. Let (X, d, T) be a topological dynamical system where X is compact and $T : X \to X$ is a continuous function which is surjective.[33] Let U be an *open cover* of X, i.e., a set $U := \{U_1, \ldots, U_k\}$, $k \in \mathbb{N}$, of open sets such that $\bigcup_{i=1}^{k} U_i \supseteq X$.[34] An open cover $V = \{V_1, \ldots, V_l\}$ is said to be an *open subcover* of an open cover U iff $V_j \in U$ for all j, $1 \leq j \leq l$. For open covers $U = \{U_1, \ldots, U_k\}$ and $V = \{V_1, \ldots, V_l\}$ let $U \vee V$ be the open cover $\{U_i \cap V_j \mid 1 \leq i \leq k;\ 1 \leq j \leq l\}$. Now for an open cover U let $N(U)$ be the minimum of the cardinality of an open subcover of U and let $h(U) := \log[N(U)]$. The following limit exists (Petersen 1983, pp. 264–265):

$$h(U, T) := \lim_{n \to \infty} \frac{h(U \vee T^{-1}(U) \vee \ldots \vee T^{-n+1}(U))}{n}, \tag{40}$$

and the topological entropy is

$$S_{top,\,A}(X, d, T) := \sup_{U} h(U, T). \tag{41}$$

$h(U, T)$ measures how the open cover U spreads out under the dynamics of the system. Hence $S_{top,\,A}(X, d, T)$ is a measure for the highest possible spreading of an open cover under the dynamics of the system. In other words, the topological entropy measures how the map T scatters states in phase space (Petersen 1983, p. 266). Note that this interpretation does not involve any probabilistic notions.

Having positive topological entropy is often linked to chaotic behaviour. For a compact phase space a positive topological entropy indicates that rel-

[33] T is required to be surjective because only then it holds that for any open cover U also $T^{-t}(U)$, $t \in \mathbb{N}$, is an open cover.

[34] Every open cover of a compact set has a finite subcover; hence we can assume that U is finite.

ative to some open covers the system scatters states in phase space. If scattering is regarded as indicating chaos, a positive entropy indicates that there is chaotic motion on some regions of the phase space. But there are many dynamical systems whose phase space is not compact; then $S_{top,\,A}(X, d, T)$ cannot be applied to distinguish chaotic from nonchaotic behaviour.

How does the topological entropy relate to the Kolmogorov-Sinai entropy? Let (X, d, T) be a topological dynamical system where X is compact and T is continuous, and denote by $M_{(X,d)}$ the set of all measure-preserving dynamical systems (X, Σ, μ, T) where Σ is the Borel σ-algebra of (X, d).[35] Then (Goodwyn 1972):

$$S_{top,\,A}(X, d, T) = \sup_{(X,\Sigma,\mu,T)\in M_{(X,d)}} S_{KS}, (X, \Sigma, \mu, T). \tag{42}$$

Furthermore, it is often said that the topological entropy is an *analogy* of the KS-entropy (e.g., Bowen 1970, p. 23; Petersen 1983, p. 264), but without providing an elaboration of the notion of analogy at work. An analogy is more than a similarity. Hesse (1963) distinguishes two kinds of analogy, material and formal. Two objects stand in material analogy, if they share certain intrinsic properties; they stand in formal analogy if they are described by the same mathematical expressions but without sharing any other intrinsic properties (see also Polya 1954). This leaves the question of what it means for *definitions* to be analogous. We say that definitions are materially/formally *analogous* iff there is a material/formal analogy between the objects appealed to in the definition.

The question then is whether $S_{top,\,A}(X, d, T)$ is analogous to the KS-entropy. Clearly, they are formally analogous: relate open covers U to partitions α, $U \vee V$ to $\alpha \vee \beta$, and $h(U)$ to $H(\alpha)$. Then, $h(U, T) = \lim_{n\to\infty}(U \vee T^{-1}(U) \ldots T^{-n+1}(U))/n$ corresponds to $H(\alpha, T) = \lim_{n\to\infty} H(\alpha \vee T^{-1}(\alpha) \ldots T^{-n+1}(\alpha))/n$, and $S_{top,\,A}(X, d, T) = \sup_U h(U, T)$ corresponds to $S_{KS}(X, \Sigma, \mu, T) = \sup_\alpha h(\alpha, T)$. However, these definitions are not materially analogous. First, $H(\alpha)$ can be interpreted as corresponding to the Shannon entropy but $h(U)$ cannot because of the absence of probabilistic notions in its definition. This seems to link it more to the Hartley entropy, which also does not explicitly appeal to probabilities: we could regard $h(U)$ as the Hartley entropy of a subcover V of U with the least elements (cf. section 3). However, this does not work because, except for the trivial open cover X, no open cover represents a set of *mutually exclusive* possibilities. Second, $h(U)$ measures the logarithm of the minimum number

[35]The Borel σ-algebra of a metric space (X, d) is the σ-algebra generated by all open sets of (X, d).

of elements of U needed to cover X, but $H(\alpha)$ has no similar interpretation, e.g., is not the logarithm of the number of elements of the partition α. Thus $S_{top,\,A}(X, d, T)$ and the KS-entropy are not materially analogous.

Bowen (1971) introduced two definitions which are equivalent to Adler et al.'s definition. Because of lack of space, we cannot discuss them here (see Petersen 1983, pp. 264–267). What matters is that there is neither a formal nor a material analogy between the Bowen entropies and the KS-entropy. Consequently, all we have is a formal analogy between the KS-entropy and the topological entropy (41), and the claims in the literature that the KS-entropy and the topological entropy are analogous are to some extent misleading. Moreover, we conclude that the topological entropy does not capture what entropy notions are often believed to capture, such as information, and that none of the interpretations of the topological entropy is similar in interpretation to another notion of entropy.

6 Fractal Geometry

It was not until the late 1960s that mathematicians and physicists started to systematically investigate irregular sets. Mandelbrot coined the term *fractal* to denote these irregular sets. Fractals have been praised for providing a better representation of several natural phenomena than figures of classical geometry but whether this is true remains controversial (cf. Falconer 1990, p. xiii; Mandelbrot 1983; Shenker 1994).

Fractal dimensions measure the irregularity of a set. We will discuss those fractal dimensions which are called *entropy dimensions*. The basic idea underlying fractal dimensions is that a set is a fractal *if* the fractal dimension is greater than the usual topological dimension (which is an integer). Yet the converse is not true: there are fractals where the relevant fractal dimensions equal the topological dimension (Falconer 1990, pp. xx-xxi and chapter 3; Mandelbrot 1983, section 39).

Fractals arise in many different contexts. In particular, in dynamical systems theory, scientists frequently focus on invariant sets, i.e., sets A for which $T_t(A) = A$ for all t, where T_t is the time evolution. And *invariant sets are often fractals*. For instance, many dynamical systems have attractors, i.e., invariant sets which neighboring states asymptotically approach in the course of dynamic evolution. Attractors are sometimes fractals, e.g., the Lorenz and the Hénon attractor.

The following idea underlies definitions of a dimension of a set F. For each $\varepsilon > 0$ we take a measurement $M_\varepsilon(F)$ of the set F at level ε, and then

we ask how $M_\varepsilon(F)$ behaves as ε goes to zero. If $M_\varepsilon(F)$ obeys the power law

$$M_\varepsilon(F) \approx c\varepsilon^{-s}, \tag{43}$$

for some constants c and s as ε goes to zero, then s is called the *dimension* of F. From (43) follows that as ε goes to zero:

$$\log[M_\varepsilon(F)] \approx \log[c] - s\log[\varepsilon]. \tag{44}$$

Consequently,

$$s = \lim_{\varepsilon\to 0} \frac{\log[M_\varepsilon(F)]}{-\log[\varepsilon]}. \tag{45}$$

If $M_\varepsilon(F)$ does not obey a power law (43), one can consider instead of the limit in (45) the limit superior and the limit inferior (cf. Falconer 1990, p. 36).

Some fractal dimensions are called entropy dimensions, namely the box-counting dimension and the Rényi entropy dimensions. Let us start with the former. Assume that $\mathbb{R}^n$ is endowed with the usual Euclidean metric d. Given a nonempty and bounded subset $F \subseteq \mathbb{R}^n$, let $B_\varepsilon(F)$ be the smallest number of balls of diameter ε that cover F. The following limit, if it exists, is called the *box-counting dimension* but is also referred to as the *entropy dimension* (Edgar 2008, p. 112; Falconer 1990, p. 38; Hawkes 1974, p. 704; Mandelbrot 1983, p. 359)

$$\mathrm{Dim}_B(F) := \lim_{\varepsilon\to 0} \frac{\log[B_\varepsilon(F)]}{-\log[\varepsilon]}. \tag{46}$$

There are several equivalent formulations of the box-counting dimension. For instance, for $\mathbb{R}^n$ consider the boxes defined by the ε-coordinate mesh:

$$[m_1\varepsilon, (m_1+1)\varepsilon) \times \ldots \times [m_n\varepsilon, (m_n+1)\varepsilon), \tag{47}$$

where $m_1, \ldots, m_n \in \mathbb{Z}$. Then if we define $B_\varepsilon(F)$ as number of boxes in the ε-coordinate mesh that intersect F, the dimension obtained is equivalent to (46) (Falconer 1990, pp. 38–39). As expected, typically, for sets of classical geometry the box dimension is integer-valued and for fractals it is non-integer valued.[36]

For instance, how many squares of side length $\varepsilon = \frac{1}{2^n}$ are needed to cover the unit square $U = [0,1] \times [0,1]$? The answer is $B_{\frac{1}{2^n}}(U) = 2^{2n}$. Hence the box-counting dimension is $\lim_{n\to\infty} \frac{\log[2^{2n}]}{-\log[\frac{1}{2^n}]} = 2$. As another example

[36]The box-counting dimension has the shortcoming that even compact countable sets can have positive dimension. Therefore, it is often modified (Edgar 2008, p. 213; Falconer 1990, p. 37 and pp. 44–46).

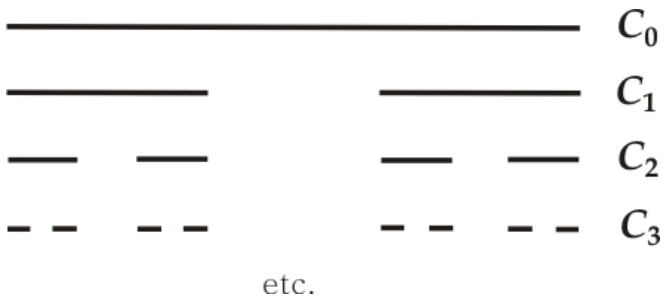

Figure 1: The Cantor Dust

we consider the *Cantor dust*, a well known fractal. Starting with the unit interval $C_0 = [0, 1]$, C_1 is obtained by removing the middle third from $[0, 1]$, C_2 is obtained by removing from C_1 the middle third of each of the intervals of C_1, and so on (see Figure 1). The Cantor dust C is defined as $\cap_{k=0}^{\infty} C_k$. By setting $\varepsilon = \frac{1}{3^n}$ and by considering the ε-coordinate mesh, we see that $B_{\frac{1}{3^n}}(C) = 2^n$. Hence

$$\mathrm{Dim}_B(C) := \lim_{n\to\infty} \frac{\log[2^n]}{-\log[\frac{1}{3^n}]} = \frac{\log[2]}{\log[3]} \approx 0.6309. \tag{48}$$

The box-counting dimension can readily be interpreted as the value of the coefficient s such that $B_\varepsilon(F)$ obeys the power law $B_\varepsilon(F) \approx c\varepsilon^{-s}$ as ε goes to zero. That is, it measures how spread out the set is when examined at an infinitesimally small scale. However, this interpretation does not link to any entropy notions. So is there such a link?

Indeed there is (surprisingly, we have been unable to identify this argument in print).[37] Consider the box-counting dimension where $B_\varepsilon(F)$ is the number of boxes in the ε-coordinate mesh that intersect F. Assume that each of these boxes represent a possible outcome and that we want to know what the actual outcome is. This assumption is sometimes natural. For instance, when we are interested in the dynamics on an invariant set F of a dynamical system we might ask: in which of the boxes of the ε-coordinate mesh that intersect F is the state of the system? Then the information gained when we learn which box the system occupies is quantified by the Hartley entropy $\log[B_\varepsilon(F)]$. Hence the box-dimension measures how the Hartley information is growing as ε goes to zero. Thus there is a link

[37]Moreover, Hawkes (1974, p. 703) refers to $\log[B_\varepsilon(F)]$ as ε-entropy, which is backed up by Kolmogorov & Tihomirov (1961) who justify calling $\log[B_\varepsilon(F)]$ entropy by an appeal to Shannon's source coding theorem. However, as they themselves observe, this justification relies on assumptions that have no relevance for scientific problems.

between the box-dimension and the Hartley entropy.

Let us now turn to the Rényi entropy dimensions. Assume that $\mathbb{R}^n$, $n \geq 1$, is endowed with the usual Euclidean metric. Let $(\mathbb{R}^n, \Sigma, \mu)$ be a measure space where Σ contains all open sets of $\mathbb{R}^n$ and where $\mu(\mathbb{R}^n) = 1$. First, we need to introduce the notion of the support of the measure μ, which is the set on which the measure is concentrated. Formally, the *support* is the smallest closed set X such that $\mu(\mathbb{R}^n \setminus X) = 0$. For instance, when measuring the dimension of a set F, the support of the measure is typically F. We assume that the support of μ is contained in a bounded region of $\mathbb{R}^n$.

Consider the ε-coordinate mesh of $\mathbb{R}^n$ (47). Let $B^i_\varepsilon, 1 \leq i \leq m$, $m \in \mathbb{N}$, be the boxes that intersect the support of μ, and let $Z_{q,\varepsilon} := \sum_{i=1}^m \mu(B^i_\varepsilon)^q$. The Rényi entropy dimension of order q, $-\infty < q < \infty$, $q \neq 1$, is

$$\mathrm{Dim}_q := \lim_{\varepsilon \to 0} \frac{1}{q-1} \frac{\log[Z_{q,\varepsilon}]}{\log[\varepsilon]}, \tag{49}$$

and the Rényi entropy dimension of order 1 is

$$\mathrm{Dim}_1 := \lim_{\varepsilon \to 0} \lim_{q \to 1} \frac{1}{q-1} \frac{\log[Z_{q,\varepsilon}]}{\log[\varepsilon]}, \tag{50}$$

if the limit exists.

It is not hard to see that if $q < q'$, $\mathrm{Dim}_{q'} \leq \mathrm{Dim}_q$ (cf. Beck & Schlögl 1995, p. 117). The cases $q = 0$ and $q = 1$ are of particular interest. Because $\mathrm{Dim}_0 = \mathrm{Dim}_B(\mathrm{support}\mu)$, the Rényi entropy dimensions are a generalisation of the box-counting dimension. And for $q = 1$ (Rényi 1961): $\mathrm{Dim}_1 = \lim_{\varepsilon \to 0} \frac{\sum_{i=1}^m -\mu(B^i_\varepsilon) \log[\mu(B^i_\varepsilon)]}{-\log(\varepsilon)}$. Since $\sum_{i=1}^m -\mu(B^i_\varepsilon) \log[\mu(B^i_\varepsilon)]$ is the *Shannon entropy* (cf. section 3), Dim_1 is called the *information dimension* (Falconer 1990, p. 260; Ott 2002, p. 81).

The Rényi entropy dimensions are often referred to as entropy dimensions (e.g., Beck & Schlögl 1995, pp. 115–116). Before turning to a rationale for this name, let us state the usual motivation of the Rényi entropy dimensions. The number q determines how much weight we assign to μ: the higher q, the greater the influence of boxes with larger measure. So the Rényi entropy dimensions measure the coefficient s such that $Z_{q,\varepsilon}$ obeys the power law $Z_{q,\varepsilon} \approx c\varepsilon^{-(1-q)s}$ as ε goes to zero. That is, Dim_q measures how spread out the support of μ is when it is examined at an infinitesimally small scale and when the weight of the measure is q (Beck & Schlögl 1995, p. 116; Ott, 2002, pp. 80–85). Consequently, when the Rényi entropy dimensions differ for different q, this is a sign of a *multifractal*, i.e., a set with different scaling behaviour (see Falconer 1990, pp. 254–264). This motivation does not refer

to entropy notions.

Yet there is an obvious connection of the Rényi entropy dimensions for $q > 0$ to the Rényi entropies (cf. section 3).[38] Assume that each of the boxes of the ε-coordinate mesh which intersect the support of μ represent a possible outcome. Further, assume that the probability that the outcome is in the box B_i is $\mu(B_i)$. Then the information gained when we learn which box the system occupies can be quantified by the Rényi entropies H_q. Consequently, each Rényi entropy dimension for $q \in (0, \infty)$ measures how the information is growing as ε goes to zero. For $q = 1$ we get a measure of how the Shannon information is growing as ε goes to zero.

7 Conclusion

This chapter has been concerned with some of the most important notions of entropy. The interpretations of these entropies have been discussed and their connections have been clarified. Two points deserve attention. First, all notions of entropy discussed in this chapter, except the thermodynamic and the topological entropy, can be understood as variants of some information theoretic notion of entropy. However, this should not distract from the fact that different notions of entropy have different meanings and play different roles. Second, there is no preferred interpretation of the probabilities that figure in the different notions of entropy. The probabilities occurring in information theoretic entropies are naturally interpreted as epistemic probabilities, but ontic probabilities are not ruled out. The probabilities in other entropies, for instance the different Boltzmann entropies, are most naturally understood ontically. So when considering the relation between entropy and probability are no simple and general answers, and a careful case by case analysis is the only way forward.

Acknowledgements

We would like to thank David Lavis, Robert Batterman and two anonymous referees for comments on earlier drafts of this chapter. We are grateful to Claus Beisbart and Stephan Hartmann for editing this book.

[38] Surprisingly, we have not found this motivation in print.

References

Adler, R., Konheim, A. & McAndrew, A. (1965), 'Topological entropy', *Transactions of the American Mathematical Society* **114**, 309–319.

Bashkirov, A.G. (2006), 'Rényi entropy as a statistical entropy for complex systems', *Theoretical and Mathematical Physics* **149**, 1559–1573.

Batterman, R. & White, H. (1996), 'Chaos and algorithmic complexity', *Foundations of Physics* **26**, 307–336.

Beck, C. & Schlögl, F. (1995), *Thermodynamics of Chaotic Systems*, Cambridge University Press, Cambridge.

Berger, A. (2001), *Chaos and Chance, an Introduction to Stochastic Aspects of Dynamics*, De Gruyter, New York.

Berkovitz, J., Frigg, R. & Kronz, F. (2006), 'The ergodic hierarchy, randomness and Hamiltonian chaos', *Studies in History and Philosophy of Modern Physics* **37**, 661–691.

Bowen, R. (1970), Topological entropy and Axiom A, *in* 'Global Analysis, Proceedings of the Symposium of Pure Mathematics 14', American Mathematical Society, Providence.

Bowen, R. (1971), 'Periodic points and measures for Axiom A diffeomorphisms', *Transactions of the American Mathematical Society* **154**, 377–397.

Cornfeld, I., Fomin, S. & Sinai, Y. (1982), *Ergodic Theory*, Springer, Berlin.

Earman, J. (1971), 'Laplacian determinism, or is this any way to run a universe?', *Journal of Philosophy* **68**, 729–744.

Eckmann, J.-P. & Ruelle, D. (1985), 'Ergodic theory of chaos and strange attractors', *Reviews of Modern Physics* **57**, 617–654.

Edgar, G. (2008), *Measure, Topology, and Fractal Geometry*, Springer, New York.

Ehrenfest, P. & Ehrenfest, T. (1912), *The Conceptual Foundations of the Statistical Approach in Mechanics*, Dover, Mineola/New York.

Emch, G. & Liu, C. (2002), *The Logic of Thermostatistical Physics*, Springer, Berlin, Heidelberg.

Falconer, K. (1990), *Fractal Geometry: Mathematical Foundations and Applications*, John Wiley & Sons, New York.

Frigg, R. (2004), 'In what sense is the Kolmogorov-Sinai entropy a measure for chaotic behaviour?—bridging the gap between dynamical systems theory and communication theory', *The British Journal for the Philosophy of Science* **55**, 411–434.

Frigg, R. (2006), 'Chaos and randomness: An equivalence proof of a generalised version of the Shannon entropy and the Kolmogorov-Sinai entropy for Hamiltonian dynamical systems', *Chaos, Solitons and Fractals* **28**, 26–31.

Frigg, R. (2008), A field guide to recent work on the foundations of statistical mechanics, *in* D. Rickles, ed., 'The Ashgate Companion to Contemporary Philosophy of Physics', Ashgate, London, pp. 99–196.

Frigg, R. (2009*a*), 'Typicality and the approach to equilibrium in Boltzmannian statistical mechanics', *Philosophy of Science (Supplement)* **76**, pp. 997–1008.

Frigg, R. (2009*b*), Why typicality does not explain the approach to equilibrium, *in* M. Suárez, ed., 'Probabilities, Causes and Propensitites in Physics', Springer, Berlin, forthcoming.

Frigg, R. & Hoefer, C. (2010), Determinism and Chance from a Humean Perspective, *in* D. Dieks, W. Gonzalez, S. Hartmann, M. Weber, F. Stadler & T. Uebel, eds., 'The Present Situation in the Philosophy of Science', Springer, Berlin, forthcoming.

Gibbs, J. (1981), *Elementary Principles in Statistical Mechanics*, Ox Bow Press, Woodbridge.

Gillies, D. (2000), *Philosophical Theories of Probability*, London: Routledge.

Goldstein, S. (2001), Boltzmann's approach to statistical mechanics, *in* J. Bricmont, D. Dürr, M. Galavotti, G. Ghirardi, F. Pettrucione & N. Zanghi, eds, 'Chance in Physics: Foundations and Perspectives', Springer, Berlin and New York, pp. 39–54.

Goodwyn, L. (1972), 'Comparing topological entropy with measure-theoretic entropy', *American Journal of Mathematics* **94**, 366–388.

Grad, H. (1961), 'The many faces of entropy', *Communications in Pure and Applied Mathematics* **14**, 323–254.

Greiner, W., Neise, L. & Stücker, H. (1993), *Thermodynamik und Statistische Mechanik*, Harri Deutsch, Leipzig.

Halmos, P. (1950), *Measure Theory*, Van Nostrand, New York and London.

Hartley, R. (1928), 'Transmission of information', *The Bell System Technical Journal* **7**, 535–563.

Hawkes, J. (1974), 'Hausdorff measure, entropy, and the independence of small sets', *Proceedings of the London Mathematical Society* **28**, 700–723.

Hemmo, M. & Shenker, O. (2006), 'Von Neumann's entropy does not correspond to thermodynamic entropy', *Philosophy of Science* **73**, 153–174.

Hesse, M. (1963), *Models and Analogies in Science*, Sheed and Ward, London.

Howson, C. (1995), 'Theories of probability', *The British Journal for the Philosophy of Science* **46**, 1-32.

Howson, C. & Urbach, P. (2006), *Scientific Reasoning: The Bayesian Approach*, Open Court, La Salle.

Ihara, S. (1993), *Information Theory for Continuous Systems*, World Scientific, London.

Jaynes, E. (1957), 'Information theory and statistical mechanics', *Physical Review* **106**, 620–630.

Jaynes, E. (1965), 'Gibbs versus Boltzmann entropies', *American Journal of Physics***33**, 391–398.

Jaynes, E. (1983), *Papers on Probability, Statistics, and Statistical Physics*, Reidel, Dordrecht.

Jizba, P. & Arimitsu, T. (2004), 'The world according to Rényi: thermodynamics of multifractal systems', *Annals of Physics* **312**, 17–59.

Kittel, C. (1958), *Elementary Statistical Mechanics*, Dover, Mineola/NY.

Klir, G. (2006), *Uncertainty and Information: Foundations of Generalized Information Theory*, Wiley, Hoboken, New Jersey.

Kolmogorov, A. (1958), 'A new metric invariant of transitive dynamical systems and automorphisms of Lebesgue spaces', *Dokl. Acad. Nauk SSSR* **119**, 861–864.

Kolmogorov, A. & Tihomirov, V. (1961), 'ε-entropy and ε-capacity of sets in functional spaces', *American Mathematical Society Translations* **17**, 277–364.

Lavis, D. (2004), 'The spin-echo system reconsidered', *Foundations of Physics* **34**, 669–688.

Lavis, D. (2008), 'Boltzmann, Gibbs, and the Concept of Equilibrium', *Philosophy of Science* **75**, 682–696.

Lebowitz, J. (1999), 'Statistical mechanics: A selective review of two central issues', *Reviews of Modern Physics* **71**, 346–357.

Lewis, D. (1980), A Subjectivists Guide to Objective Chance, *in* R.C. Jeffrey, ed., 'Studies in Inductive Logic and Probability', University of California Press, Berkeley, pp. 83–132.

Mañé, R. (1987), *Ergodic Theory and Differentiable Dynamics*, Springer, Berlin.

Mandelbrot, B. (1983), *The Fractal Geometry of Nature*, Freeman, New York.

Mellor, H. (2005), *Probability: A Philosophical Introduction*, Routledge, London.

Ott, E. (2002), *Chaos in Dynamical Systems*, Cambridge University Press, Cambridge.

Penrose, R. (1970), *Foundations of Statistical Mechanics*, Oxford University Press, Oxford.

Petersen, K. (1983), *Ergodic Theory*, Cambridge University Press, Cambridge.

Pippard, A.B. (1966), *The Elements of Classical Thermodynamics.*, Cambridge University Press, Cambridge.

Polya, G. (1954), *Patterns of Plausible Inference*, Volume II of Mathematics and Plausible Reasoning, Princeton University Press, Princeton.

Reiss, H. (1965), *Methods of Thermodynamics*, Dover Publications, Mineola/NY.

Rényi, A. (1961), On measures of entropy and information, *in* 'Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability', University of California Press, Berkeley, pp. 547–561.

Shannon, C. & Weaver, W. (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Chicago, IL & London.

Shaw, R. (1985), *The Dripping Faucet as a Model Chaotic System*, Aerial Press, Santa Cruz.

Shenker, O. (1994), 'Fractal geometry is not the geometry of nature', *Studies in History and Philosophy of Modern Physics* **25**, 967–981.

Sinai, Y. (1959), 'On the concept of entropy for dynamical systems', *Dokl. Acad. Nauk SSSR* **124**, 768–771.

Sklar, L. (1993), *Physics and Chance. Philosophical Issues in the Foundations of Statistical Mechanics*, Cambridge University Press, Cambridge.

Sorkin, R. (2005), 'Ten theses on black hole entropy', *Studies In History and Philosophy of Modern Physics* **36**, 291–301.

Sutherland, W. (2002), *Introduction to Metric and Topological Spaces*, Oxford University Press, Oxford.

Tolman, R. (1938), *The Principles of Statistical Mechanics*, Dover, Mineola and New York.

Tsallis, C. (1988), 'Possible generalization of Boltzmann-Gibbs statistics', *Journal of Statistical Physics* **52**, 479–487.

Uffink, J. (2001), 'Bluff your way in the second law of thermodynamics', *Studies in the History and Philosophy of Modern Physics* **32**, 305–394.

Uffink, J. (2006), Compendium to the foundations of classical statistical physics, *in* J. Butterfield & J. Earman, eds, 'Philosophy of Physics', North-Holland, Amsterdam, pp. 923–1074.

Wehrl, A. (1978), 'General properties of entropy', *Reviews of Modern Physics* **50**, 221–259.

Werndl, C. (2009*a*), 'Are deterministic descriptions and indeterministic descriptions observationally equivalent?', *Studies in History and Philosophy of Modern Physics* **40**, 232–242.

Werndl, C. (2009*b*), Deterministic versus indeterministic descriptions: Not that different after all?, *in* A. Hieke & H. Leitgeb, eds, 'Reduction, Abstraction, Analysis, Proceedings of the 31st International Ludwig Wittgenstein-Symposium', Ontos, Frankfurt, pp. 63–78.

Werndl, C. (2009*c*), 'Justifying definitions in matemathics—going beyond Lakatos', *Philosophia Mathematica* **17**, pp. 313–340.

Werndl, C. (2009*d*), 'What are the new implications of chaos for unpredictability?', *The British Journal for the Philosophy of Science* **60**, 195–220.