Entropy **2008**, *10*, 493-506; DOI: 10.3390/e10040493

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Entropy and Uncertainty

Derek W. Robinson

Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia
E-mail: Derek.Robinson@anu.edu.au

Received: 17 June 2008 / Accepted: 6 August 2008 / Published: 16 October 2008

Abstract: We give a survey of the basic statistical ideas underlying the definition of entropy in information theory and their connections with the entropy in the theory of dynamical systems and in statistical mechanics.

Keywords: Entropy, relative entropy, uncertainty, information theory.

---

1. Introduction

The concept of entropy originated in the physical and engineering sciences but now plays a ubiquitous role in all areas of science and in many non-scientific disciplines. A quick search of the ANU library catalogue gives books on entropy in mathematics, physics, chemistry, biology, communication theory and engineering but also in economics, linguistics, music, architecture, urban planning, social and cultural theory and even in creationism. Many of the scientific applications will be described in the lectures over the next three weeks. In this brief introductory lecture we describe some of the theoretical ideas that underpin these applications.

Entropy is an encapsulation of the rather nebulous notions of disorder or chaos, uncertainty or randomness. It was introduced by Clausius in the 19th century in thermodynamics and was an integral part of Boltzmann's theory. In the thermodynamic context the emphasis was on entropy as a measure of disorder. Subsequently the probabilistic nature of the concept emerged more clearly with Gibbs work on statistical mechanics. Entropy then had a major renaissance in the middle of the 20th century with the development of Shannon's mathematical theory of communication. It was one of Shannon's great insights that entropy could be used as a measure of information content. There is some ambiguity in the notion of information and in Shannon's theory it is less a measure of what one communicates but rather what one could communicate, i.e. it is a measure of one's freedom of choice when one selects a

message to be communicated. Shannon also stressed the importance of the relative entropy as a measure of redundancy. The relative entropy gives a comparison between two probabilistic systems and typically measures the actual entropy to the maximal possible entropy. It is the relative entropy that has played the key role in many of the later developments and applications.

Another major landmark in the mathematical theory of the entropy was the construction by Kolmogorov and Sinai of an isomorphy invariant for dynamical systems. The invariant corresponds to a mean value over time of the entropy of the system. It was remarkable as it differed in character from all previous spectral invariants and it provided a mechanism for providing a complete classification of some important systems. Other landmarks were the definition of mean entropy as an affine functional over the state space of operator algebras describing models of statistical mechanics and the application of this functional to the characterization of equilibrium states. Subsequently entropy became a useful concept in the classification of operator algebras independently of any physical background.

In the sequel we discuss entropy, relative entropy and conditional entropy in the general framework of probability theory. In this context the entropy is best interpreted as a measure of uncertainty. Subsequently we develop some applications, give some simple examples and indicate how the theory of entropy has been extended to non-commutative settings such as quantum mechanics and operator algebras.

2. Entropy and uncertainty

First consider a probabilistic process with n possible outcomes. If $n = 2$ this might be something as simple as the toss of a coin or something as complex as a federal election. Fortunately the details of the process are unimportant for the sequel. Initially we make the simplifying assumption that all n outcomes are equally probable, i.e. each outcome has probability $p = 1/n$. It is clear that the inherent uncertainty of the system, the uncertainty that a specified outcome actually occurs, is an increasing function of n. Let us denote the value of this function by $f(n)$. It is equally clear that $f(1) = 0$ since there is no uncertainty if there is only one possible outcome. Moreover, if one considers two independent systems with n and m outcomes, respectively, then the combined system has nm possible outcomes and one would expect the uncertainty to be the sum of the individual uncertainties. In symbols one would expect

$$f(nm) = f(n) + f(m) \ .$$

The additivity reflects the independence of the two processes. But if this property is valid for all positive integers n, m then it is easy to deduce that

$$f(n) = \log n$$

although the base of logarithms remains arbitrary. (It is natural to chose base 2 as this ensures that $f(2) = 1$, i.e. a system such as coin tossing is defined to have unit uncertainty, but other choices might be more convenient.) Thus the uncertainty per outcome is given by $(1/n) \log n$ or, expressed in terms of the probability,

$$\text{uncertainty per outcome} = -p \log p \ .$$

Secondly, consider a process with n possible outcomes with probabilities $p_1, p_2, \ldots, p_n$, respectively.

Then it is natural to ascribe uncertainty $-p_i \log p_i$ to the i-th outcome. This is consistent with the foregoing discussion and leads to the hypothesis

$$\text{total uncertainty} = -\sum_{i=1}^{n} p_i \log p_i \ .$$

This is the standard expression for the entropy of the probabilistic process and we will denote it by the symbol $H(p)$, or $H(p_1, \ldots, p_n)$. (This choice of notation dates back to Boltzmann.) Explicitly, the **entropy** is defined by

$$H(p) = H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \ . \tag{1}$$

Although the argument we have given to identify the entropy with the inherent uncertainty is of a rather *ad hoc* nature applications establish that it gives a surprisingly efficient description. This will be illustrated in the sequel. It is a case of 'the proof of the pudding is in the eating'.

Before proceeding we note that the entropy $H(p)$ satisfies two simple bounds. If $x \in \langle 0, 1]$ then $-x \log x \in [0, 1]$ and one can extend the function $x \mapsto -x \log x$ to the closed interval $[0, 1]$ by setting $-0 \log 0 = 0$. Then $H(p) \geq 0$ with equality if and only if one outcome has probability one and the others have probability zero. It is also straightforward to deduce that $H(p)$ is maximal if and only if the probabilities are all equal, i.e. if and only if $p_1 = \ldots = p_n = 1/n$. Therefore one has bounds

$$0 \leq H(p_1, \ldots, p_n) \leq \log n \ . \tag{2}$$

There is a third less precise principle: the most probable outcomes give the major contribution to the total entropy.

The entropy enters the Boltzmann–Gibbs description of equilibrium statistical mechanics through the prescription that the state of equilibrium is given by the microscopic particle configurations which maximize the entropy under the constraints imposed by the observation of macroscopic quantities such as the energy and density. Thus if the configuration with assigned probability p_i has energy e_i the idea is to maximize $H(p)$ with $E(p) = \sum_{i=1}^{n} p_i e_i$ held fixed. If one formulates this problem in terms of a Lagrange multipliers $\beta \in \mathbf{R}$ then one must maximize the function

$$p = (p_1, \ldots, p_n) \mapsto H(p) - \beta\, E(p) \ . \tag{3}$$

We will discuss this problem later in the lecture.

3. Entropy and multinomial coefficients

In applications to areas such as statistical mechanics one is usually dealing with systems with a large number of possible configurations. The significance of entropy is that it governs the asymptotic behaviour of the multinomial coefficients

$${}^{n}C_{n_1 \ldots n_m} = \frac{n!}{n_1! \ldots n_m!}$$

where $n_1 + \ldots + n_m = n$. These coefficients express the number of ways one can divide n objects into m subsets of $n_1, n_2, \ldots, n_m$ objects, respectively. If n is large then it is natural to examine the number

of partitions into m subsets with fixed proportions $p_1 = n_1/n, \ldots, p_m = n_m/n$. Thus one examines

$$P_n(p) = \frac{n!}{(p_1 n)! \ldots (p_m n)!}$$

with $p_1 + \ldots + p_n = 1$ as $n \to \infty$. Since the sum over all possible partitions,

$$\sum_{n_1,\ldots,n_m} {}^n C_{n_1 \ldots n_m} = m^n \ ,$$

increases exponentially with n one expects a similar behaviour for $P_n(p)$. Therefore the asymptotic behaviour will be governed by the function $n^{-1} \log P_n(p)$. But this is easily estimated by use of the Stirling-type bounds

$$(2\pi n)^{1/2} n^n e^{-n} e^{1/(12n+1)} \leq n! \leq (2\pi n)^{1/2} n^n e^{-n} e^{1/(12n)}$$

for the factorials. One finds

$$n^{-1} \log P_n(p) = H(p) + O(-n^{-1} \log n^{-1}) \tag{4}$$

as $n \to \infty$ where $H(p)$ is the entropy. Thus the predominant asymptotic feature of he P_n is an exponential increase $\exp(nH)$ with H the entropy of the partition $p_1, \ldots, p_m$.

Next consider n independent repetitions of an experiment with m possible outcomes and corresponding probabilities $q_1, \ldots, q_m$. The probability that these outcomes occur with frequencies $p_1, \ldots, p_m$ is given by

$$P_n(p|q) = P_n(p)\, q_1^{p_1 n} \ldots q_m^{p_m n} = \frac{n!}{(p_1 n)! \ldots (p_m n)!}\, q_1^{p_1 n} \ldots q_m^{p_m n} \ .$$

Then estimating as before one finds

$$n^{-1} \log P_n(p|q) = H(p|q) + O(-n^{-1} \log n^{-1}) \tag{5}$$

as $n \to \infty$ where $H(p|q)$ is given by

$$H(p|q) = -\sum_{i=1}^{m} (p_i \log p_i - p_i \log q_i) \ . \tag{6}$$

The latter expression is the **relative entropy** of the frequencies p_i with respect to the probabilities q_i. Now it is readily established that $H(p|q) \leq 0$ with equality if, and only if, $p_i = q_i$ of all $i \in \{1, \ldots, m\}$. Thus if the p_i and q_i are not equal then $P_n(p|q)$ decreases exponentially as $n \to \infty$. Therefore the only results which effectively occur are those for which the frequencies closely approximate the probabilities.

One can relate the variational principle (3) defining the Boltzmann–Gibbs equilibrium states to the relative entropy. Set $q_i = e^{-\beta e_i}/Z$ where $Z = \sum_i e^{-\beta e_i}$. Then

$$H(p) - \beta\, E(p) = H(p|q) + \log Z \ .$$

Therefore the maximum is obtained for $p_i = q_i$ and the maximal value is $\log Z$.

4. Conditional entropy and information

Next we consider two processes α and β and introduce the notation $A_1, \ldots, A_n$ for the possible outcomes of α and $B_1, \ldots, B_m$ for the possible outcomes of β. The corresponding probabilities are denoted by $p(A_1), \ldots, p(A_n)$ and $p(B_1), \ldots, p(B_m)$, respectively. The joint process, α followed by β, is denoted by $\alpha\beta$ and the probability of the outcome A_iB_j is given by $p(A_iB_j)$. We assume that $p(A_iB_j) = p(B_jA_i)$ although this condition has to be relaxed in the subsequent non-commutative settings. If the α and β are independent processes then $p(A_iB_j) = p(A_i)p(B_j)$ and the identity is automatic.

The probability that B_j occurs given the prior knowledge that A_i has occurred is called the **conditional probability** for B_j given A_i. It is denoted by $p(B_j|A_i)$. A moment's reflection establishes that

$$p(A_i)p(B_j|A_i) = p(A_iB_j) \ .$$

Since the possible outcomes B_jA_i of the process $\beta\alpha$ are the same as the outcomes A_iB_j of the process $\alpha\beta$ one has the relation

$$p(A_i)p(B_j|A_i) = p(B_j)p(A_i|B_j)$$

for the conditional probabilities. If the two processes are independent one obviously has $p(B_j|A_i) = p(B_j)$.

The conditional entropy of the process β given the outcome A_i for α is then defined by

$$H(\beta|A_i) = -\sum_{j=1}^{m} p(B_j|A_i) \log p(B_j|A_i) \tag{7}$$

in direct analogy with the (unconditional) entropy of β defined by

$$H(\beta) = -\sum_{j=1}^{m} p(B_j) \log p(B_j) \ .$$

In fact if α and β are independent then $H(\beta|A_i) = H(\beta)$. Since A_i occurs with probability $p(A_i)$ it is natural to define the **conditional entropy** of β dependent on α by

$$H(\beta|\alpha) = \sum_{i=1}^{n} p(A_i) H(\beta|A_i) \ . \tag{8}$$

The entropy $H(\beta)$ is interpreted as the uncertainty in the process β and $H(\beta|\alpha)$ is the residual uncertainty after the process α has occurred.

Based on the foregoing intuition Shannon defined the difference

$$I(\beta|\alpha) = H(\beta) - H(\beta|\alpha) \tag{9}$$

as the **information** about β gained by knowledge of the outcome of α. It corresponds to the reduction in uncertainty of the process β arising from knowledge of α.

The usefulness of these various concepts depends on a range of properties that are all traced back to simple features of the function $x \mapsto -x \log x$.

First, one has the key relation

$$H(\beta|\alpha) = H(\alpha\beta) - H(\alpha) \ . \tag{10}$$

This follows by calculating that

$$\begin{aligned} H(\beta|\alpha) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_iB_j)\log p(B_j|A_i) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_iB_j)\log(p(A_iB_j)/p(A_i)) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_iB_j)\left(\log p(A_iB_j) - \log p(A_i)\right) = H(\alpha\beta) - H(\alpha) \end{aligned}$$

Note that if α and β are independent then $H(\beta|\alpha) = H(\beta)$ and the relation (10) asserts that $H(\alpha\beta) = H(\alpha) + H(\beta)$. Note also that it follows from (10) that the information is given by

$$I(\beta|\alpha) = H(\beta) + H(\alpha) - H(\alpha\beta)\ . \tag{11}$$

This latter identity establishes the symmetry

$$I(\beta|\alpha) = I(\alpha|\beta)\ . \tag{12}$$

Next remark that $H(\beta|\alpha) \geq 0$ by definition. Hence $H(\alpha\beta) \geq H(\alpha)$ and, by symmetry, $H(\alpha\beta) \geq H(\beta)$. Thus one has the lower bounds

$$H(\alpha\beta) \geq H(\alpha) \vee H(\beta)\ . \tag{13}$$

But it also follows by a convexity argument that $H(\beta|\alpha) \leq H(\beta)$. The argument is as follows. Since the function $x \geq 0 \mapsto \log x$ is convex one has

$$\sum_{i=1}^{n} \lambda_i \log x_i \leq \log(\sum_{i=1}^{n} \lambda_i x_i)$$

for all $\lambda_i \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$ and all $x_i \geq 0$. Therefore

$$\begin{aligned} H(\beta|\alpha) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(A_iB_j)\log p(B_j|A_i) \\ &= \sum_{j=1}^{m} p(B_j)\sum_{i=1}^{n}(p(A_iB_j)/p(B_j))\log(1/p(B_j|A_i)) \\ &\leq \sum_{j=1}^{m} p(B_j)\log\sum_{i=1}^{n}(p(A_i)/p(B_j)) = H(\beta) \end{aligned}$$

because $\sum_{i=1}^{n}(p(A_iB_j)/p(B_j)) = 1$. (Here I have been rather cavalier in assuming $p(B_j|A_i)$ and $p(B_j)$ are strictly positive but it is not difficult to fill in the details.)

It follows from $H(\beta|\alpha) \leq H(\beta)$ that

$$I(\beta|\alpha) \geq 0\ , \tag{14}$$

i.e. the information is positive. But then using the identity (11) one deduces that

$$H(\alpha\beta) \leq H(\alpha) + H(\beta)\ . \tag{15}$$

This is a generalization of the property of subadditivity $f(x+y) \leq f(x) + f(y)$ for functions of a real variable. It is subsequently of fundamental importance.

Finally we note that the information is increasing in the sense that

$$I(\beta\gamma|\alpha) \geq I(\beta|\alpha) \tag{16}$$

or, equivalently,

$$H(\alpha\beta\gamma) - H(\beta\gamma) \leq H(\alpha\beta) - H(\beta)\ . \tag{17}$$

The latter property is established by calculating that

$$\begin{aligned} H(\alpha\beta\gamma) - H(\alpha\beta) - H(\beta\gamma) + H(\beta) &= -\sum_{i,j,k} p(A_iB_jC_k) \log \frac{p(A_iB_jC_k)p(B_j)}{p(A_iB_j)p(B_jC_k)} \\ &\leq \sum_{i,j,k} \frac{p(A_iB_j)p(B_jC_k)}{p(B_j)} \Big(1 - \frac{p(A_iB_jC_k)p(B_j)}{p(A_iB_j)p(B_jC_k)}\Big) \\ &= \sum_{i,j,k} \Big(\frac{p(A_iB_j)p(B_jC_k)}{p(B_j)} - p(A_iB_jC_k)\Big) = 0 \end{aligned}$$

where we have used the bound $-x \log x \leq 1 - x$. This property (17) is usually referred to as strong subadditivity as it reduces to the subadditive condition (15) if β is the trivial process with only one outcome.

Example There are two cities, for example Melbourne and Canberra, and the citizens of one always tells the truth but the citizens of the other never tell the truth. An absent-minded mathematician forgets where he is and attempts to find out by asking a passerby, who could be from either city. What is the least number of questions he must ask if the only replies are 'yes' and 'no'? Alternatively, how many questions must he pose to find out where he is and where the passerby lives?

Since there are two towns there are two possible outcomes to the experiment α of questioning. If the mathematician really has no idea where he is then the entropy $H(\alpha) = \log 2$ represents the total information. Then if one uses base 2 logarithms $H(\alpha) = 1$. So the problem is to ask a question β that gives unit information, i.e. such that $I(\alpha, \beta) = H(\alpha) = 1$ or, equivalently, $H_\beta(\alpha) = 0$. Thus the question must be unconditional. This could be achieved by asking 'Do you live here?'.

Alternatively to find out where he is and to also decide where the passerby lives he needs to resolve the outcome of a joint experiment $\alpha_1\alpha_2$ where α_1 consists of finding his own location and α_2 consists of finding the residence of the passerby. But then the total information is $H(\alpha_1\alpha_2) = H(\alpha_1) + H_{\alpha_1}(\alpha_2) > 1$. Hence one question will no longer suffice but two clearly do suffice: he can find out where he is with one question and then find out where the passerby lives with a second question. This is consistent with the fact that $H(\alpha_1\alpha_2) = \log 2^2 = 2$.

This is all rather easy. The following problem is slightly more complicated but can be resolved by similar reasoning.

Problem Assume in addition there is a third city, Sydney say, in which the inhabitants alternately tell the truth or tell a lie. Argue that the mathematician can find out where he is with two questions but needs four questions to find out in addition where the passerby lives.

5. Dynamical systems

Let $(X, \mathcal{B})$ denote a σ-finite measure space, i.e. a set X equipped with a σ-algebra $\mathcal{B}$ of subsets of X. Further let μ denote a probability measure on $(X, \mathcal{B})$. Then $(X, \mathcal{B}, \mu)$ is called a probability space. A finite partition $\alpha = (A_1, \ldots, A_n)$ of the space is a collection of a finite number of disjoint elements A_i of $\mathcal{B}$ such that $\bigcup_{i=1}^n A_i = X$. Given two partitions $\alpha = (A_1, \ldots, A_n)$ and $\beta = (B_1, \ldots, B_m)$ the join $\alpha \vee \beta$ is defined as the partition composed of the subsets $A_i \cap B_j$ with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$.

If $\alpha = (A_1, \ldots, A_n)$ is a partition of the space then $0 \leq \mu(A_i) \leq 1$ and $\sum_{i=1}^n \mu(A_i) = 1$ because μ is a probability measure. Thus the $\mu(A_i)$ correspond to the probabilities introduced earlier and $\alpha \vee \beta$ now corresponds to the joint process $\alpha\beta$. Therefore we can now use the definitions of entropy, conditional entropy, etc. introduced previously but with the replacements $p(A_i) \to \mu(A_i)$, $\alpha\beta \to \alpha \vee \beta$ etc. Note that since $\mu(A_i \cap B_j) = \mu(B_j \cap A_i)$ one automatically has $p(A_i B_j) = p(B_j A_i)$.

Next let T be a measure-preserving invertible transformation of the probability space $(X, \mathcal{B}, \mu)$. In particular $T\mathcal{B} = \mathcal{B}$ and $\mu(TA) = \mu(A)$ for all $A \in \mathcal{B}$. Then $(X, \mathcal{B}, \mu, T)$ is called a dynamical system. In applications one also encounters dynamical systems in which T is replaced by a measure-preserving flow, i.e. a one-parameter family of measure preserving transformations $\{T_t\}_{t \in \mathbf{R}}$ such that $T_s T_t = T_{s+t}$, $T_{-s} = (T_s)^{-1}$ and $T_0 = I$ is the identity. The flow is usually interpreted as a description of the change with time t of the observables A. The single automorphism T can be thought of as the change with unit time and T^n is the change after n-units of time.

The entropy of the partition $\alpha = (A_1, \ldots, A_n)$ of the probability space is given by

$$H(\alpha; \mu) = -\sum_{i=1}^n \mu(A_i) \log \mu(A_i)$$

and we now define the **mean entropy** of the partition of the dynamical system by

$$H(\alpha\,; \mu, T) = \lim_{n \to \infty} n^{-1} H(\alpha \vee T\alpha \ldots \vee T^n \alpha; \mu) \tag{18}$$

where $T\alpha = (TA_1, \ldots, TA_n)$. Then the mean entropy of the automorphism T is defined by

$$H(\mu, T) = \sup_\alpha H(\alpha\,; \mu, T) \tag{19}$$

where the supremum is over all finite partitions.

It is of course necessary to establish that the limit in (18) exists. But this is a consequence of subadditivity. Set

$$f(n) = H(\alpha \vee T\alpha \ldots \vee T^{n-1}\alpha; \mu)$$

then

$$\begin{aligned} f(n+m) &= H((\alpha \vee T\alpha \ldots \vee T^{n-1}\alpha; \mu) \vee T^n(\alpha \vee T\alpha \ldots \vee T^{m-1}\alpha; \mu) \\ &\leq H(\alpha \vee T\alpha \ldots \vee T^{n-1}\alpha; \mu) + H(T^n(\alpha \vee T\alpha \ldots \vee T^{m-1}\alpha); \mu) = f(n) + f(m) \end{aligned}$$

for all $n, m \in \mathbf{N}$ where we have used (15) and the T-invariance of μ. It is, however, an easy consequence of subadditivity that the limit of $n^{-1} f(n)$ exists as $n \to \infty$.

The entropy $H(\mu, T)$ was introduced by Kolmogorov and Sinai and is often referred to as the Kolmogorov–Sinai invariant. This terminology arises since $H(\mu, T)$ is an isomorphy invariant. Two dynamical systems $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ are defined to be isomorphic if there is an invertible measure preserving map U of $(X_1, \mathcal{B}_1, \mu_1)$ onto $(X_2, \mathcal{B}_2, \mu_2)$ which intertwines T_1 and T_2, i.e. which has the property $UT_1 = T_2U$. If this is the case then $H(\mu_1, T_1) = H(\mu_2, T_2)$. Thus to show that two dynamical systems are not isomorphic it suffices to prove that $H(\mu_1, T_1) \neq H(\mu_2, T_2)$. Of course this is not necessarily easy since it requires calculating the entropies. This is, however, facilitated by a result of Kolmogorov and Sinai which establishes that the supremum in (19) is attained for a special class of partitions. The partition α is defined to be a **generator** if $\bigvee_{k=-\infty}^{\infty} T^k\alpha$ is the partition of X into points. Then

$$H(\mu, T) = H(\alpha\,; \mu, T) \tag{20}$$

for each finite generator α.

Another way of formulating the isomorphy property is in Hilbert space terms. Let $\mathcal{H} = L_2(X\,; \mu)$. Then it follows from the T-invariance property of μ that there is a unitary operator on $\mathcal{H}$ such that $Uf = f \circ T^{-1}$ for all $f \in C_b(X)$. Now the two dynamical systems $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ are isomorphic if there is a unitary operator V from $\mathcal{H}_1$ to $\mathcal{H}_2$ which intertwines the two unitary representatives U_1 and U_2 of the maps T_1 and T_2. Since one has $U_1 = VU_2V^{-1}$ the spectra of U_1 and U_2 are also isomorphy invariants. The Kolgmogorov–Sinai entropy was the first invariant which was not of a spectral nature.

The mean entropy has another interesting property as a function of the measure μ. The probability measures form a convex subset of the dual of the bounded continuous functions $C_b(X)$. Now if μ_1 and μ_2 are two probability measures and $\lambda \in [0, 1]$ then

$$H(\alpha; \lambda\mu_1 + (1-\lambda)\mu_2) \geq \lambda\, H(\alpha; \mu_1) + (1-\lambda)\, H(\alpha; \mu_2)\ . \tag{21}$$

The concavity inequality (21) is a direct consequence of the definition of $H(\alpha; \mu)$ and the concavity of the function $x \mapsto -x \log x$. Conversely, one has inequalities

$$-\log(\lambda\, \mu_1(A_i) + (1-\lambda)\, \mu_2(A_i)) \leq -\log\lambda - \log\mu_1(A_i)$$

and

$$-\log(\lambda\, \mu_1(A_i) + (1-\lambda)\, \mu_2(A_i)) \leq -\log(1-\lambda) - \log\mu_2(A_i)$$

because $x \mapsto -\log x$ is decreasing. Therefore one obtains the 'convexity' bound

$$H(\alpha; \lambda\mu_1 + (1-\lambda)\mu_2) \leq \lambda\, H(\alpha; \mu_1) + (1-\lambda)\, H(\alpha; \mu_2) - \lambda\log\lambda - (1-\lambda)\log(1-\lambda)\ . \tag{22}$$

Now replacing α by $\alpha \vee T\alpha \ldots \vee T^{n-1}\alpha$ in (21), dividing by n and taking the limit $n \to \infty$ gives

$$H(\alpha\,; \lambda\mu_1 + (1-\lambda)\mu_2, T) \geq \lambda\, H(\alpha\,; \mu_1, T) + (1-\lambda)\, H(\alpha\,; \mu_2, T)\ .$$

Similarly from (22), since $-(\lambda\log\lambda + (1-\lambda)\log(1-\lambda))/n \to 0$ as $n \to \infty$, one deduces the converse inequality

$$H(\alpha\,; \lambda\mu_1 + (1-\lambda)\mu_2, T) \leq \lambda\, H(\alpha\,; \mu_1, T) + (1-\lambda)\, H(\alpha\,; \mu_2, T)\ .$$

Hence one concludes that the map $\mu \mapsto H(\alpha\,;\mu,T)$ is affine, i.e.

$$H(\alpha\,;\lambda\mu_1+(1-\lambda)\mu_2,T)=\lambda\,H(\alpha\,;\mu_1,T)+(1-\lambda)\,H(\alpha\,;\mu_2,T) \tag{23}$$

for each partition α, each pair μ_1 and μ_2 of probability measures and each $\lambda\in[0,1]$.

Finally it follows from the identification (20) that the mean entropy is also affine,

$$H(\lambda\mu_1+(1-\lambda)\mu_2,T)=\lambda\,H(\mu_1,T)+(1-\lambda)\,H(\mu_2,T) \tag{24}$$

for each pair μ_1 and μ_2 of probability measures and each $\lambda\in[0,1]$. This is a somewhat surprising and is of great significance in the application of mean entropy in statistical mechanics.

6. Mean entropy and statistical mechanics

The simplest model of statistical mechanics is the one-dimensional ferromagnetic Ising model. This describes atoms at the points of a one-dimensional lattice $\mathbf{Z}$ with two degrees of freedom which we label as $0,1$ and which we think of as a spin orientation. Thus $X=\{0,1\}^{\mathbf{Z}}$ and a point $x\in X$ is a doubly-infinite array of 0s and 1s. The labels indicate whether a particle at a given point of the lattice has negative or positive spin orientation. Two neighbouring atoms with identical orientation are ascribed a negative unit energy and neighbouring atoms with opposite orientation are ascribed a positive unit of energy. Therefore it is energetically favourable for the atoms to align and provide a spontaneous magnetism. Since the configurations x of particles are doubly infinite the total energy ascribed to each x is usually infinite but the mean energy, i.e. the energy per lattice site is always finite.

The model generalizes to d dimensions in an obvious way. Then $X=\{0,1\}^{\mathbf{Z}^d}$ and a point $x\in X$ corresponds to a d-dimensional array of 0s and 1s. If one now assigns a negative unit energy to each pair of nearest neighbours in the lattice $\mathbf{Z}^d$ with similar orientations and positive unit energy to the pairs with opposite orientation then the energy of a configuration of particles on a cubic subset of $\mathbf{Z}^d$ with side length L grows as L^d, i.e. as the d-dimensional volume. Therefore the energy per lattice site is again finite.

The group $\mathbf{Z}^d$ of shifts acts in an obvious manner on X. Let $T_1,\ldots,T_d$ denote the unit shift to in each of the d directions. Further let μ denote a $\mathbf{Z}^d$-invariant probability measure over X. Then the energy $E(\mu)$ per lattice site is well defined, it has a value in $[-1,1]$ and $\mu\mapsto E(\mu)$ is an affine function. Now consider the entropy per lattice site.

Let α denote the partition of X into two subsets, the subset A_0 of configurations with a 0 at the origin of $\mathbf{Z}^d$ and the subset A_1 of configurations with a 1 at the origin. Then $\bigvee_{k_1=-\infty}^{\infty}\cdots\bigvee_{k_d=-\infty}^{\infty}T_1^{k_1}\ldots T_d^{k_d}=X$ and the partition α is a generator of X. Now the previous definition of the mean entropy generalizes and

$$H(\mu)=\lim_{L_1,\ldots,L_d\to\infty}(L_1\ldots L_d)^{-1}H(\bigvee_{k_1=0}^{L_1}\ldots\bigvee_{k_d=0}^{L_d}T_1^{k_1}\ldots T_d^{k_d}\alpha:\mu) \tag{25}$$

exists by an extension of the earlier subadditivity argument to the d-dimensional setting.

The Boltzmann–Gibbs approach sketched earlier would designate the equilibrium state of the system at fixed mean energy as the measure which maximizes the functional

$$\mu\mapsto H(\mu)-\beta\,E(\mu)\ .$$

This resembles the earlier algorithm but there is one vital difference. Now the supremum is taken over the infinite family of invariant probability measures μ over X. There is no reason that the supremum is uniquely attained. In fact this is not usually the case.

There is a competition between two effects. Assuming $\beta > 0$ the energy term $-\beta E(\mu)$ is larger if $E(\mu)$ is negative and this requires alignment of the spins, i.e. ordered configurations are preferred. But the entropy term $H(\mu)$ is largest if the the system is disordered, i.e. if all possible configurations are equally possible. If β is large the energy term tends to prevail but if β is small then the entropy term prevails. In fact β is interpretable as the inverse temperature and there is a tendency to ordering at low temperatures and to disorder at high temperatures. Since there are two possible directions of alignment of the spins this indicates that there are two distinct maximising measures at low temperature and only one at high temperatures. The advantage of this description is that it reflects reality. The Ising model, with $d \geq 2$, indeed gives a simple description of a phase transition for which there is a spontaneous magnetization at low temperatures.

Although we have described the model with a nearest neighbour interaction which favours alignment of the model atoms the same general features pertain if the interaction favours anti-parallel alignment, i.e. if the alignment of neighbours has positive energy and the anti-alignment negative energy. Then it is still energetically favourable to have an ordered state but the type of ordering is different. The model then describes a phenomenon called anti-ferrogmagnetism.

The description of the invariant equilibrium states as the invariant measures which maximize the mean entropy at fixed mean energy has many other positive aspects. Since $\mu \mapsto H(\mu) - \beta\, E(\mu)$ is an affine function it tends to attain its maximum at extremal points of the convex weakly* compact set of invariant measures E. In fact if the maximum is unique then the maximizing measure is automatically extremal. If, however, the maximum is not uniquely attained then the maximizing measures form a face Δ_β of the convex set E and each $\mu \in \Delta_\beta$ has a unique decomposition as a convex combination of extremal measures in Δ_β. This indicates that the extremal measures correspond to pure phases and in the case of a phase transition there is a unique prescription of the phase separation. This interpretation is corroborated by the observation that the extremal invariant states are characterized by the absence of long range correlations.

The foregoing description of the thermodynamic phases of macroscopic systems was successfully developed in the 1970s and 1980s and also extended to the description of quantum systems. But the latter extension requires the development of a non-commutative generalization of the entropy.

7. Quantum mechanics and non-commutativity

The Ising model has a simple quantum-mechanical extension. Again one envisages atoms at the points of a cubic lattice $\mathbf{Z}^d$ but each atom now has more structure. The simplest assumption is that the observables corresponding to the atom at the point $x \in \mathbf{Z}^d$ are described by an algebra $\mathcal{A}_{\{x\}}$ of 2×2-matrices. Then the observables corresponding to the atoms at the points of a finite subset $\Lambda \subset \mathbf{Z}^d$ are described by an algebra $\mathcal{A}_\Lambda$ of $2^{|\Lambda|} \times 2^{|\Lambda|}$-matrices where $|\Lambda|$ indicates the number of points in Λ. Thus

$$\mathcal{A}_\Lambda = \prod\nolimits_{x\in\Lambda}^{\otimes} \mathcal{A}_{\{x\}}$$

where the product is a tensor product of matrices. A quantum-mechanical state ω_λ of the subsystem Λ is then determined by a positive matrix ρ_Λ with $\mathrm{Tr}_\Lambda(\rho_\Lambda) = 1$ where Tr_Λ denotes the trace over the matrices $\mathcal{A}_\Lambda$. The value of an observable $A \in \mathcal{A}_\Lambda$ in the state ω_λ is then given by

$$\omega_\Lambda(A) = \mathrm{Tr}_\Lambda(\rho_\Lambda A)$$

Now if $\Lambda \subset \Lambda'$ one can identify $\mathcal{A}_\Lambda$ as a subalgebra of $\mathcal{A}_{\Lambda'}$ and for consistency the matrices ρ_Λ that determine the state must satisfy the condition

$$\rho_\Lambda = \mathrm{Tr}_{\Lambda'\setminus\Lambda}(\rho_{\Lambda'}) \ . \tag{26}$$

The natural generalization of the classical entropy is now given by the family of entropies

$$H_\Lambda(\omega) = -\,\mathrm{Tr}_\Lambda(\rho_\Lambda \log \rho_\Lambda) \tag{27}$$

as Λ varies over the bounded subsets of $\mathbf{Z}^d$. The previous mean entropy should then be defined by

$$H(\omega) = \lim_{\Lambda\to\mathbf{Z}^d} H_\Lambda(\omega)/|\Lambda|$$

if the limit exists. The existence of the limit is now a rather different problem than before. Nevertheless it can be established for translation invariant states by a extension of the earlier subadditivity argument which we now briefly describe.

First if ρ and σ are two positive matrices both with unit trace then the entropy of ρ relative to σ is defined by

$$H(\rho|\sigma) = -\,\mathrm{Tr}(\rho\log\rho - \rho\log\sigma)$$

in direct analogy with the earlier definition (6). The key point is that one still has the property $H(\rho|\sigma) \leq 0$. This is established as follows. Let ρ_i and σ_i denote the eigenvalues of ρ and σ. Further let ψ_i denote an orthonormal family of eigenfunctions of ρ corresponding to the eigenvalues ρ_i. Then

$$\begin{aligned}
-\,\mathrm{Tr}(\rho\log\rho - \rho\log\sigma) &= -\sum_i(\rho_i\log\rho_i - \rho_i(\psi_i, \log\sigma\psi_i)) \\
&\leq -\sum_i(\rho_i\log\rho_i - \rho_i\log(\psi_i, \sigma\psi_i)) \\
&= -\sum_i(\psi_i, \sigma\psi_i)(\rho_i/(\psi_i, \sigma\psi_i))\log(\rho_i/(\psi_i, \sigma\psi_i)) \\
&\leq \sum_i(\psi_i, \sigma\psi_i)(1 - \rho_i/(\psi_i, \sigma\psi_i)) = 1 - 1 = 0
\end{aligned}$$

where we have used convexity of the logarithm and the inequality $-x\log x \leq 1 - x$.

Now suppose that Λ_1 and Λ_2 are two disjoint subsets of $\mathbf{Z}^d$. Set $\rho = \rho_{\Lambda_1\cup\Lambda_2}$ and $\sigma = \rho_{\Lambda_1}\otimes\rho_{\Lambda_2}$. Then it follows from the foregoing that

$$-\,\mathrm{Tr}_{\Lambda_1\cup\Lambda_2}\left(\rho_{\Lambda_1\cup\Lambda_2}\log\rho_{\Lambda_1\cup\Lambda_2} - \rho_{\Lambda_1\cup\Lambda_2}\log(\rho_{\Lambda_1}\otimes\rho_{\Lambda_2})\right) \leq 0 \ .$$

But using (26) and the identity

$$\log(\rho_{\Lambda_1}\otimes\rho_{\Lambda_2}) = \log(\rho_{\Lambda_1})\otimes\mathbb{1}_{\Lambda_2} + \mathbb{1}_{\Lambda_1}\otimes\log(\rho_{\Lambda_2})$$

one immediately deduces that

$$H_{\Lambda_1 \cup \Lambda_2}(\omega) \leq H_{\Lambda_1}(\omega) + H_{\Lambda_2}(\omega) \ .$$

This corresponds to the earlier subadditivity and suffices to prove the existence of the mean entropy.

These simple observations on matrix algebras are the starting point of the development of a non-commutative entropy theory.

Bibliography

There is an enormous literature on entropy but the 1948 paper *A mathematical theory of communication* by Claude Shannon in the Bell System Technical Journal remains one of the most readable accounts of its properties and significance [1]. This paper can be downloaded from

http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

Note that in later versions it became *The mathematical theory of communication*.

Another highly readable account of the entropy and its applications to communication theory is given in the book *Probability and Information* by A. M. Yaglom and I. M. Yaglom. This book was first published in Russian in 1956 as *Veronatjost' i informacija* and republished in 1959 and 1973. It has also been translated into French, German, English and several other languages. The French version was published by Dunod and an expanded English version was published in 1983 by the Hindustan Publishing Corporation [2]. I was unable to find any downloadable version.

The example of the absentminded mathematician was adapted from this book which contains many more recreational examples and a variety of applications to language, music, genetics etc.

The discussion of the asymptotics of the multinomial coefficients is apocryphal. I have followed the discussion in the Notes and Remarks to Chapter VI of *Operator Algebras and Quantum Statistical Mechanics* 2 by Bratteli and Robinson, Springer-Verlag, 1981 [3]. This book is now available as two searchable pdf files:

http://folk.uio.no/bratteli/bratrob/VOL-1S 1.pdf
http://folk.uio.no/bratteli/bratrob/VOL-2.pdf.

There are now many books that describe the ergodic theory of dynamical systems including the theory of entropy. The earliest source in English which covered the Kolmogorov–Sinai theory was I believe the 1962–63 Aarhus lecture notes of Jacobs *Ergodic theory* I and II. These are now difficult to find but are worth reading if you can locate a copy. Another early source is the book by Arnold and Avez *Problèmes ergodiques de la mecanique classique*, Gauthier-Villars, Paris 1967 [4].

More recent books which I have found useful are *An Introduction to Ergodic Theory* by Ya. G. Sinai, Princeton University Press, Mathematical Notes, 1976 [5]: *An Introduction to Ergodic Theory* by P. Walters, Springer-Verlag, Graduate Text in Mathematics, 1981 [6]: *Topics in Ergodic Theory* by W. Parry, Cambridge University Press, 1981 [7]. But there are many more.

Chapter VI of *Operator Algebras and Quantum Statistical Mechanics* 2 by Bratteli and Robinson [3] contains a description of the applications of entropy to spin systems but the theory has moved on since then. Another source which covers more recent developments in the quantum-mechanical applications

is *Quantum entropy and its use* by M. Ohya and D. Petz, Springer-Verlag, 1993 [8]. Finally a recent comprehensive treatment of the extension of the theory to the structural study of operator algebras is given in *Dynamical Entropy in Operator Algebras* by S. Nesheyev and E. Størmer, Springer-Verlag, 2006 [9].

References and Notes

1. Shannon, C. A mathematical theory of communication. *Bell Sys. Tech.* **1948**, *27*, 379-423.
2. Yaglom, A. M.; Yaglom, I. M. In *Probability and information* ; Hindustan Publishing Corporation: India, 1983.
3. Bratteli, O.; Robinson, D.W. In *Operator algebras and quantum statistical mechanics 2* ; Springer-Verlag: New York, 1981.
4. Arnold, V.I. ; Avez, A. In *Problèmes ergodiques de la mecanique classique* ; Gauthier-Villars: Paris, 1967.
5. Sinai, Ya. G. *An introduction to ergodic theory* ; Princeton University Press: New Jersey, 1976.
6. Walters, P. *An introduction to ergodic theory* ; Springer-Verlag: New York, 1981.
7. Parry, W. *Topics in ergodic theory* ; Cambridge University Press: UK, 1981.
8. Ohya, M.; Petz, D. In *Quantum entropy and its use* ; Springer-Verlag: Berlin, 1993.
9. Nesheyev, S.; Strmer, E. In *Dynamical entropy in operator algebras* ; Springer-Verlag: Berlin, 2006.