Information and Entropy

J.M. Schumacher
CWI,
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands
Tilburg University, Dept. of Econometrics,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

The related notions of 'information' and 'entropy' are basic concepts in statistics as well as in theoretical physics and in theoretical computer science. The present paper provides a survey of the origin of these notions in the mathematical theory of communication as developed by Claude Shannon, and of their use in the three mentioned fields. The paper is the written version of a talk held in one of a series of 'reconstruction seminars' organized by a group of system theorists in the Netherlands. To ensure the introductory level of the lectures in these seminars, all presentations are given by non-experts.

1. The reconstruction seminar

This paper[1] is the written version of a presentation held in January 1993 before a group of system theorists. The presentation was part of a series of so-called 'reconstruction seminars'. To explain the background of the present paper, it is necessary to say a bit more about this seminar series. In September 1991, a number of system theorists working in the Netherlands received a letter from Jan Willems, Christiaan Heij, and Paula Rocha. In their letter, the threesome expressed their dissatisfaction with the overly specialized and technical nature of many conference talks, and with the lack of opportunity to discuss in some depth developments outside of the immediate research environment. They proposed to organize a series of 'reconstruction seminars' which would serve as sort of an antidote to this state of affairs. The idea would be to form a group of researchers in system theory, each of whom would be assigned a subject outside his or her own field of expertise, in order to present a lecture on that subject before the group. The time available for the lecture and the discussion would be three hours per subject. In this way, the group would broaden its knowledge of a variety of fields with a limited time investment.

[1] I would like to thank the participants in the reconstruction seminar for their unsparing criticism of the original presentation. Also, I would like to thank Peter Gács for his comments on a preliminary version of this paper. Of course, any remaining errors are entirely my own responsibility.

This idea fell into fertile ground; after a preparation period in which subjects were selected and assigned, the first seminar series took place in the spring semester of 1993. There were fourteen subjects in this first series, and as many speakers. The participants were allowed to bring one guest with them and so the group actually attending the seminars consisted of about twenty-five people. The subjects in the first series would be hard to fit into one category (they included wavelets, coding theory, quantum mechanics, inverse scattering, and cointegration), but they are all in some way highlights in the general area of engineering, physics, econometrics and mathematics. Discussions have certainly been livelier than they are at most conferences, and the learning rate was also higher. A second series of reconstruction seminars is planned for 1995.

To avoid any possible misunderstanding that might arise in connection with the term 'system theory', let me point out the following. Mathematical system theory is a branch of applied mathematics dealing with the relations between time-dependent variables as described by systems of differential and algebraic equations or by other means. It is true that the field draws upon a large variety of mathematical disciplines and has applications in many branches of engineering, and this may be one of reasons why the idea of the reconstruction seminar was well-received; but in no way is there a pretension in the seminar that all subjects treated could be brought under a unifying framework, such as may be the belief of proponents of a 'general system theory'. The system theorists taking part in the seminar are down-to-earth applied mathematicians who certainly have an interest in powerful general settings, but who don't like to waste their time on over-generalization.

It should perhaps also be explained where the name 'reconstruction seminar' comes from. Jan Willems told the following story about the origin of the name. At some time during 1991, the three initiators were having a drink in a *Heurigen* (a place where new wine is served) with a Soviet scientist who claimed that he had just 'reconstructed' himself. It turned out that he intended this term to be a translation of the Russian word 'perestroika'. The accuracy of this translation may be doubted—'restructuring' is perhaps closer to the original meaning of the word than 'reconstruction'—but nevertheless, or maybe just for that reason, Jan and his companions thought that the name 'reconstruction seminar' would be appropriate for what they were contemplating at the time.

As mentioned above, the present paper is the written version of one of the lectures in the first series of the reconstruction seminar. The author certainly satisfies the requirement of being a non-expert in his subject. The presentation that is to follow will therefore be introductory and relatively non-technical (as intended in the reconstruction seminar); completeness is not an objective. Actually the subject of 'information and entropy' is a vast one, providing ample room for many lifetime scientific careers. It could seem preposterous to write a survey of such a subject. Nevertheless, if one believes that there is a continuum of possibilities between knowing nothing and knowing everything about a subject, then it should also be possible to say something sensible on a given topic within such an arbitrary time span as three hours, for an audience with a decent general background in mathematics but little or no specific knowledge of the subject. The present paper is the result of an attempt to do just this.

2. The mathematical theory of communication

'Information' is a word from everyday life, but is also used in a more technical sense, sometimes with the related notion of 'entropy', in various disciplines such as communications engineering, statistics, computer science, and physics, in particular thermodynamics and statistical mechanics. The roots of the term 'entropy' lie in nineteenth-century physics, but information theory as such is usually considered as a product of the period following the Second World War, and more in particular as the brain child of Claude E. Shannon. It seems fitting, therefore, to begin a survey of information and entropy with a description of Shannon's original development of the theory.

Claude Elwood Shannon (1916–) worked as a communications engineer for the Bell Company during the Second World War, where he was involved in the coding of messages such as the ones that were exchanged between Roosevelt and Churchill. Most likely this is where he got the ideas about the upper limits of efficiency of coding that led him to his 'mathematical theory of communication'. Shannon's original paper 'A mathematical theory of communication' was published in the *Bell System Technical Journal* in two installments in 1948 [14]. The paper was soon reprinted in a booklet published by the University of Illinois Press [15] under the title *The Mathematical Theory of Communication*, which also included a 'non-technical' introduction by Warren Weaver. The slight change of title expressed perhaps already growing confidence that indeed a new branch of science had been born.

It would certainly seem ambitious to develop a mathematical theory of communication as this concept is understood in everyday life. As happens more often, the only way to say much about the subject is to discuss just a part of it. In the introduction to his paper, Shannon immediately stresses that he will exclude the *meaning* of messages from his theory. His argument is simple and effective: the semantic aspects are irrelevant to the engineer who is faced with the problem of designing a communication system.

> The significant aspect is that the actual message is one *selected from a set* [Shannon's emphasis] of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. [15, p. 31]

The term 'each possible selection' here is somewhat crude, since the important aspect which forms the basis of Shannon's theory is the *statistical* nature of messages to be sent. The basic observation is that the efficiency of coding can be improved by making use of information on the relative frequency of symbols. Given a source which emits symbols (or rather a statistical model of that source, which assigns probabilities to strings of symbols), there is an upper bound on the efficiency of coding schemes for that source. This upper bound can be expressed in standard units (say, average number of bits per symbol) and so it gives a number that quantifies a rather basic property of the source. That property is what Shannon calls the *entropy*. Given this idea, the obvious problem is how to *compute* the entropy for a given statistical model of a source, and that is the main subject of Shannon's paper.

Such may have been the line of thought that Shannon followed, but it is not the line of his exposition. In his paper he first introduces the entropy in an axiomatic way, and only later he shows how the entropy is related to the maximal efficiency of coding. His starting point is the following question: how can we define one number $H(p_1, \cdots, p_n)$ that represents the 'uncertainty' in a situation in which n possible events may occur with probabilities p_1, $\cdots$, p_n. The function H should satisfy the following requirements:

(i) H is continuous in the p_i's;

(ii) for $p_i = \frac{1}{n}$, H should be increasing as a function of n;

(iii) 'weighted additivity'[2] holds.

Shannon proves that any function that satisfies the above requirements is necessarily of the form[3]

$$H(p_1, \cdots, p_n) \;=\; -k \sum_{i=1}^{n} p_i \log p_i$$

where k is a positive constant. The p_i's all lie between 0 and 1 so that the expressions $p_i \log p_i$ are all negative or zero; therefore, the minus sign in the above expression serves to make $H(p_1, \cdots, p_n)$ nonnegative for all possible values of the p_i's. Since furthermore the numbers p_i must add up to one, the possibility $H(p_1, \cdots, p_n) = 0$ occurs only when one of the p_i's is equal to 1 and all the others are zero, which indeed corresponds to a situation of 'zero uncertainty'.

The formula above contains a constant k which remains to be specified; equivalently, one could change the base of the logarithm. The choice of k corresponds to a choice of 'unit of information'. If one takes logarithms to the base 2 and fixes k by requiring that

$$H(\tfrac{1}{2}, \tfrac{1}{2}) = 1,$$

the corresponding unit of information is called the 'binary digit' or *bit*. The abbreviation, which now belongs to the vocabulary of elementary school students,

[2] The meaning of this can be explained in an example. Suppose that event A occurs with probability $\frac{1}{2}$, and that if A does not occur, either B or C occurs with probability $\frac{2}{3}$ and $\frac{1}{3}$ respectively. In total, the events A, B, and C occur with probabilities $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{6}$ respectively. Weighted additivity requires that $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2} H(\frac{2}{3}, \frac{1}{3})$, where the factor $\frac{1}{2}$ appears because the choice between B and C only has to be made in one half of the cases.

[3] By convention, the value of $x \log x$ at 0 is taken to be 0.

was dubbed by John Tukey of Bell Laboratories [5]. When writing his paper, Shannon apparently didn't consider the term 'bit' to be firmly entrenched, for he frequently uses the longer phrase instead.

The quantity $H(p_1, \cdots, p_n)$ is called the *entropy* of the probability distribution given by the numbers $p_1, \cdots, p_n$. Shannon borrowed this term from physics, and in his paper he briefly refers to the use of the notion of entropy in statistical mechanics, mentioning Tolman's book *Principles of Statistical Mechanics*. The standard anecdote on how Shannon got to use this term is the following one (reported in *Scientific American*, 1971). Shannon recalls his deliberations on how to name the quantity that he had defined:

> My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, You should call it entropy, for two reasons. In the first place, your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.

We shall come back to the use of the word 'entropy' in physics later on in this paper. Let it just be mentioned here that Von Neumann's suggestion stirred considerable discussion about the relation between entropy in information theory and entropy in physics.

If one considers the information generated by a single event, it is quite natural to use the measure $I(p) = -\log p$, where p denotes the probability of the event. Such a measure was in fact suggested by Wiener in 1948 [18]. It is a standard fact in analysis that the logarithmic function is the only one that satisfies $I(pq) = I(p) + I(q)$ for all p and q in the interval $(0, 1]$; in other words, the logarithmic information measure is the only one such that the amount of information in two independent events is the sum of the amounts of information in the two events separately. The choice of base 2 for the logarithm is enforced by the normalization $I(\frac{1}{2}) = 1$. Shannon's entropy $H(p_1, \cdots, p_n)$ can then be interpreted as the *expected amount of information* from an experiment that has n possible outcomes with known probabilities $p_1, \cdots, p_n$. In other words, if you have a choice between several experiments, there are good reasons to choose the one that maximizes the entropy; children who play 'guess-my-number' use this fact.

2.1. Optimal coding

To get closer to the intended application, optimal coding of messages, let us now consider the entropy of an *information source*, that is, a source producing symbols from a finite alphabet according to some statistically described rule. Shannon considers in particular finite-state Markov chains such as the one depicted in Figure 1. The nodes labeled 1, 2, and 3 are called *states*. The arrows denote transitions, occurring with the indicated *transition probabilities*.

FIGURE 1. Finite-state Markov chain

The probabilities for transitions emanating from a given node must add up to 1. Each transition produces a symbol from some alphabet, in this case $\{A, B, C\}$. This model is capable of describing strings of symbols with (certain types of) statistical dependencies, like in English the letter n is more likely to precede an e than a z. The Markov chain is considered as operating for an indefinitely long time. Under mild conditions, all states i have an *equilibrium probability* P_i associated to them, which gives the average fraction of time that the chain will spend in state i, regardless of which initial state was chosen. Each state also has an entropy H_i associated to it, corresponding to the probability distribution of the symbols that may be emitted by a transition from state i. Shannon now defines the *entropy per symbol* as

$$H = \sum_i P_i H_i.$$

This weighted average characterizes the expected rate at which information is generated by the source. To add support to this interpretation, Shannon proves the following remarkable theorem. Choose any number q strictly between 0 and 1, and denote by $n(q, N)$ the minimal number of different sequences of length N emitted by the source that together have a probability at least equal to q. In other words, $\log n(q, N)$ is the number of bits that would be needed to specify $q \cdot 100\%$ of the occurring sequences of length N. Shannon shows that

$$\lim_{N \to \infty} \frac{\log n(q, N)}{N} = H$$

regardless of which q was chosen. This means that an arbitrarily large fraction of long sequences generated by the source can be specified using H bits per symbol.

The entropy of an information source can also be characterized directly from the statistical properties of the generated sequences. Shannon gives the following formula:

$$H = \lim_{N \to \infty} -\frac{1}{N} \sum_{B_i} p(B_i) \log p(B_i),$$

where the summation extends over all sequences B_i of length N. In fact the above expression could well serve as the *definition* of 'entropy per symbol'; it certainly seems natural enough, and the definition would then not be limited to sources modelled by finite-state Markov chains. For more general sources, however, it is not always clear how to compute the above limit, and this may have been the reason for Shannon to choose a more restrictive context.[4]

The first main application of the entropy concept is the celebrated *Shannon coding theorem for a noiseless channel*. The theorem can be phrased as follows.

> Consider a source with entropy H (bits/symbol) and a channel with capacity C (bits/second). For every $\varepsilon > 0$ it is possible to code the source output in such a way that an average of more than $\frac{C}{H} - \varepsilon$ symbols per second is transmitted through the channel. On the other hand, an average better than $\frac{C}{H}$ bits per second cannot be attained.

To make this more concrete, let us consider a simple example (taken from [1]). Suppose we have a source that emits either a 1 or a 0 every unit of time; a 0 occurs with probability $\frac{1}{4}$ and a 1 occurs with probability $\frac{3}{4}$, independently of the value of other symbols in the sequence. The entropy is $-0.25 \log 0.25 - 0.75 \log 0.75 = 0.81$, and so according to Shannon's theorem it should be possible to achieve an average transmission speed of one symbol per time unit even if only a channel having a capacity of 0.85 bits per time unit is available. The required compression can in this case be achieved very simply, namely by taking the emitted symbols together in groups of two. According to the rules of the source, the probability of occurrence is $\frac{1}{16}$ for 00, $\frac{3}{16}$ for 01 and 10, and $\frac{9}{16}$ for 11. If we code 11 by 1, 01 by 01, 10 by 001, and 00 by 000,[5] then the average number of bits needed to code a group of two symbols from the source is

$$\frac{9}{16} \cdot 1 + \frac{3}{16} \cdot 2 + \frac{3}{16} \cdot 3 + \frac{1}{16} \cdot 3 = \frac{27}{16} = 1.6875.$$

Dividing by 2, we obtain slightly over 0.84 for the average number of bits per symbol—sufficient to get the required transmission speed. By forming blocks of three symbols one can achieve an average number of 0.82 bits per symbol, and the entropy is reached asymptotically by taking larger and larger blocks.

In general, one gets a near-optimal coding system by looking at the probabilities of sequences of length N, where N is large, and assigning short codewords to sequences of high probability and long codewords to sequences of low probability. More precisely, the optimal codeword length for a sequence of probability

[4] The equilibrium probabilities for a finite-state Markov chain can be found by solving a simple eigenvalue problem, so the expression that Shannon uses as the definition of the entropy per symbol is indeed readily computable.

[5] This is a special case of a procedure for assigning codewords to groups of symbols known as *Huffman coding*. The coded messages will be uniquely decipherable even when the code words are concatenated, because we are using a so-called *prefix code* (cf. note 14).

p is $-\log p$ bits. It is clear (for instance from the fact that $-\log p$ is in general not an integer) that there are some problems of implementation associated with this prescription, and that the optimum can in general only be reached in the limit. As Shannon notes himself, highly efficient codes tend to require long memories and so may not be very practical. Efficiency of coding (in connection with other factors which will partly be discussed below) remains an active research field.

2.2. Mutual entropy

Many interesting and important aspects are added to the theory when we consider *mutual entropies*, which are designed to measure the information contained in one variable concerning another variable. Shannon discusses this subject extensively in his paper; his motivation is the noisy channel, in which the symbol received contains probabilistic rather than absolute information about the symbol that was sent. The theory can be formulated, however, in a rather general setting in which one speaks just about two stochastic variables x and y. The main new concepts are the *conditional entropy of* y *given* x, and the *information in* x *about* y. These concepts are obtained as follows.

Let x and y be discrete stochastic variables, each with a finite number of possible values. The two variables are in general correlated, so that the uncertainty about y will be affected if we know the outcome of x. Specifically, if we know that the outcome of x is i, then the resulting entropy of y is expressed in terms of the conditional probabilities of the outcomes of y:

$$H(y \mid x=i) \; = \; -\sum_j P(y=j \mid x=i) \log P(y=j \mid x=i).$$

Taking the weighted average with respect to the various possible outcomes of x, one obtains the conditional entropy[6] of y given x:

$$\begin{aligned} H(y \mid x) &= \sum_i P(x=i) H(y \mid x=i) \\ &= -\sum_i P(x=i) \sum_j P(y=j \mid x=i) \log P(y=j \mid x=i). \end{aligned}$$

Using Bayes' rule,

$$P(y=j \mid x=i) \; = \; \frac{P(y=j \wedge x=i)}{P(x=i)},$$

one can rewrite the conditional probabilities that appear here in terms of joint probabilities. This leads to the following expression for the conditional entropy:

$$H(y \mid x) \; = \; \sum_i \sum_j P(x=i \wedge y=j) \log \frac{P(x=i \wedge y=j)}{P(x=i)}.$$

[6] The notation used here differs from Shannon's.

One may also define a joint entropy of the two variables x and y, which is just the entropy of the variable (x, y):

$$H(x,y) \;=\; \sum_i \sum_j P(x=i \wedge y=j) \log P(x=i \wedge y=j).$$

If we now compute the difference $H(x,y) - H(y \mid x)$, we get

$$\begin{aligned} H(x,y) - H(y \mid x) &= -\sum_i \sum_j P(x=i \wedge y=j) \log P(x=i) \\ &= -\sum_i P(x=i) \log P(x=i) \\ &= H(x). \end{aligned}$$

The resulting formula

$$H(x,y) \;=\; H(x) + H(y \mid x) \tag{1}$$

may be phrased as: the uncertainty about x and y together is equal to the uncertainty about x plus the uncertainty about y when x is given. That is certainly as it should be.

Another expected property is that the conditional entropy of y given x is always less than the entropy of y, except when y and x are independent, in which case the entropies should be the same. This follows from the equality we just derived, together with the inequality $H(x,y) \leq H(x) + H(y)$ which results from[7]

$$\begin{aligned} H(x,y) &= -\sum_i \sum_j P(x=i \wedge y=j) \log P(x=i \wedge y=j) \\ &\leq -\sum_i \sum_j P(x=i \wedge y=j) \log (P(x=i) P(y=j)) \\ &= H(x) + H(y). \end{aligned}$$

Equality holds if and only if x and y are independent.[8]

Because $H(x) + H(y \mid x) = H(x,y) = H(y) + H(x \mid y)$, we can define a quantity that is symmetric in x and y by

$$I(x:y) \;=\; H(x) - H(x \mid y) \;=\; H(y) - H(y \mid x).$$

This quantity is called the 'information in x about y' (or, of course, the information in y about x). It is not difficult to get an expression for this which clearly shows the symmetry between x and y:

$$H(x) - H(x \mid y) \;=\; -\sum_i \sum_j P(x=i \wedge y=j) \log P(x=i)$$

[7] The inequality between the first and the second line is a special case of a rule which will be shown later—see (2).

[8] For this see also (2).

$$
\begin{aligned}
& + \sum_i \sum_j P(x = i \wedge y = j) \log \frac{P(x = i \wedge y = j)}{P(y = j)} \\
= \quad & \sum_i \sum_j P(x = i \wedge y = j) \log \frac{P(x = i \wedge y = j)}{P(x = i) P(y = j)}.
\end{aligned}
$$

Now, let x denote the input into a channel, and let y be the corresponding output that is received at the other end, and that may have been corrupted by noise[9]. The *capacity* of the channel is defined by Shannon as the information in y about x when the source is suitably adapted to the channel, so

$$C \;=\; \max(H(y) - H(y \mid x))$$

where the maximum is taken over all sources. This poses a maximization problem which, depending on the statistics of the channel, may or may not be difficult to solve. Shannon shows that the channel capacity is equal to the limit expression

$$C \;=\; \lim_{T \to \infty} \frac{\log N(T, q)}{T},$$

where $N(T, q)$ denotes the maximal number of sequences of length T that can be sent through the channel with probability at most q of incorrect interpretation. The formula holds for any q strictly between 1 and 0.

Shannon's *coding theorem for a noisy channel* now states the following.

> Consider a source with entropy H, and a channel with capacity C, both measured in bits per second[10]. It is possible to send information from the source through the channel with arbitrarily small error frequency by suitable encoding if and only if $H \leq C$.

Again, Shannon readily admits that the coding schemes that he presents for the sufficiency part of his proof are impractical. Error-correcting codes have become the subject of a rich field of research, with applications ranging from satellite communications to CD players.

2.3. Continuous probability distributions

Shannon devoted a large part of his paper to the study of entropy for *continuous* distributions. Here we shall only mention the most basic points. In view of the definition of entropy for discrete variables, the definition

$$H(x) \;=\; -\int p(x) \log p(x)\, dx$$

for a continuously distributed variable with density $p(x)$ strongly suggests itself. However, there is a feature here which does not appear in the discrete case.

[9] The term 'channel' refers here to the complete trajectory between sender and receiver, which includes not only the actual transmission but also all processing that is done locally.

[10] So it is assumed here that the source is emitting symbols at a fixed rate.

For discrete variables, one has $H(y) = H(x)$ whenever $y = \phi(x)$, where ϕ is an invertible mapping. If we try the effect of a transformation in the continuous-time case, however, we get (J is the Jacobian matrix of the transformation):

$$\begin{aligned} H(y) &= -\int p(y) \log p(y) \, dy \\ &= -\int p(x) \det J(x) \log(p(x) \det J(x)) \, dy \\ &= -\int p(x) [\log p(x) + \log \det J(x)] \, dx \\ &= H(x) - \int p(x) \log \det J(x) \, dx. \end{aligned}$$

The formula shows that, unless somehow a standard volume element is given, the entropy of a continuous variable is only determined up to an additive constant. Or in other words: in the continuous case, only *differences* of entropies are invariant under coordinate transformations.

It is of interest to look for distributions which maximize the entropy in a given situation; such distributions describe the 'least informative' cases, or, from another point of view, the cases in which we gain most from doing experiments. For discrete variables, the answer is the obvious one: $H(p_1, \cdots, p_n)$ is maximal when all p_i's are equal to $\frac{1}{n}$, so when all possible outcomes are equally probable. Some more calculation is needed for continuous distributions. Consider for instance the problem of determining the maximum-entropy distribution on the real line with expectation 0 and fixed variance σ^2. This leads to the following optimization problem: maximize

$$-\int p(x) \log p(x) \, dx$$

under

$$\int x^2 p(x) \, dx \; = \; \sigma^2$$

and

$$\int p(x) \, dx \; = \; 1.$$

Introducing Lagrange multipliers in the integrand, we find that

$$-p \log p + \lambda p x^2 + \mu p$$

is maximized over p when

$$-1 - \log p + \lambda x^2 + \mu = 0.$$

This determines the form of p as a function of x:

$$p(x) \; = \; \alpha e^{\beta x^2}.$$

As always in Lagrange optimization, the constants must be determined such that $p(x)$ satisfies the side constraints; this gives

$$p(x) \;=\; \frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{1}{2\sigma^2}x^2),$$

the Gaussian distribution. The conclusion can be phrased as follows: among all experiments that deliver a real value with a given variance σ^2, the ones whose outcomes are normally distributed are the most informative.

3. Entropy in physics

As noted above, the term 'entropy' had been in use long before John von Neumann suggested to Shannon to employ this word. The concept was introduced by Clausius in 1865 as a macroscopic quantity whose existence could be derived from the second law of thermodynamics. The statistical interpretation of entropy in terms of microscopic states was developed later, by Boltzmann in 1877.

On the meaning of the terms 'thermodynamics' and 'statistical mechanics', let me quote Pais [12, p. 81]:

> Such concepts as temperature, pressure, along with energy, and what came to be called entropy, are all macroscopic. The branch of science dealing with the interrelations of these quantities is called thermodynamics. It is not an easy subject. The connection between these macroscopic concepts and the underlying microscopic, molecular properties is called statistical mechanics. It is a difficult subject.

The *first law* of thermodynamics is the law of conservation of energy. The *second law* is sometimes given as 'entropy is nondecreasing', but this is not the original formulation and in fact the existence of an entropy function is derived, as will be shown below, from a principle that is stated in terms of heat, work, and temperature. One formulation, ascribed to Carnot and also to Kelvin and Planck, is the following:

> no process is possible which has as its only result the transformation of heat into work.

Another formulation, attributed to Clausius, reads as follows:

> no process is possible which has as its only result the transfer of heat from a cooler to a warmer body.

To prove that these formulations are indeed equivalent, it suffices to show that a process can be created that transfers heat from a cooler to a warmer body if a process is available that transforms heat into work, and vice versa; this is not too difficult.

FIGURE 2. Carnot cycle.

The existence of a macroscopic quantity called the entropy is derived from one of the two equivalent formulations by a reasoning which depends on the notion of *reversible processes*. Such processes are idealizations of actual processes that are nearly reversible. Using such idealizations in an early stage is perhaps typical for physical reasoning; a mathematician would probably prefer to work with processes that are as near to being reversible as one wants, and pass to the limit only later. The following two examples of irreducible processes are the standard ones, and are used in in the derivation of the entropy.

EXAMPLE 1. (isothermic expansion/compression). Consider a cylinder filled with an ideal gas, in contact with a heat reservoir of temperature T_1. Pull out a frictionless piston 'very slowly'. The gas will expand while staying at the same temperature, and meanwhile will take up heat from the reservoir. This is (in the limit) a reversible process; by pushing in the piston again 'very slowly', the whole system can be brought back into exactly the same condition as it was before the expansion.

EXAMPLE 2. (adiabatic expansion/compression). Consider the same cylinder as above, but now in thermal isolation. Again pull out the piston very slowly. The temperature of the gas will fall. This is again (in the limit) a reversible process.

A concatenation of reversible processes is of course again reversible. By combining the two types of processes just mentioned, one can obtain a reversible *cycle*: first isothermic expansion, then adiabatic expansion, then isothermic compression, and finally adiabatic compression to the original state. This is called a *Carnot cycle*; the whole process can be plotted in the (V, P)-plane as shown in Figure 2. The curves corresponding to isothermic expansion and compression satisfy $PV =$ constant (actually $PV = RT$, where T is the temperature), and the curves corresponding to adiabatic expansion and compression satisfy $PV^\gamma =$ constant, where $\gamma > 1$ is a constant which is typical of the gas used. In terms of heat and work, what happens during a Carnot cycle is that the gas in the cylinder takes up an amount of heat $(\Delta Q)_1$ from a reservoir at temperature T_1, and delivers an amount of heat $(\Delta Q)_2$ to another reservoir at

FIGURE 3. Heat engine.

a lower temperature T_2; in the whole cycle, an amount of work is done equal to $\int P\,dV$, which is exactly the surface enclosed in Figure 2. A Carnot cycle is therefore an example of a *heat engine*, the general form of which is sketched in Figure 3.[11] According to the first law of thermodynamics, the amount of work done by a heat engine must be equal to the difference between the heat extracted from the warm reservoir and the heat delivered at the cold reservoir, so

$$\Delta U \;=\; (\Delta Q)_1 + (\Delta Q)_2$$

where $(\Delta Q)_1$ is negative. One may expect, however, that a further relation can be derived between $(\Delta Q)_1$, $(\Delta Q)_2$, T_1, and T_2 from the supposition that the process is reversible. A few computations bear this out. Denote by a, b, c, d the four intermediate stages in the Carnot cycle (see Fig. 2). We have

$$(\Delta Q)_1 \;=\; \int_a^b P\,dV \;=\; \int_a^b RT_1 \frac{dV}{V} \;=\; RT_1 \log \frac{V_b}{V_a}$$

and likewise

$$(\Delta Q)_2 \;=\; RT_2 \log \frac{V_d}{V_c}.$$

Furthermore, it follows from the relation $P_b V_b^\gamma \;=\; P_c V_c^\gamma$ and the relations $P_b V_b = RT_1$ and $P_c V_c = RT_2$ that

$$T_1 V_b^{\gamma-1} \;=\; T_2 V_c^{\gamma-1}$$

and in the same way

$$T_1 V_a^{\gamma-1} \;=\; T_2 V_d^{\gamma-1}.$$

Dividing, we get $V_b/V_a = V_c/V_d$ and so from the above:

$$\frac{(\Delta Q)_1}{T_1} + \frac{(\Delta Q)_2}{T_2} \;=\; 0.$$

[11] For the purpose of the reasoning to follow, one might just as well consider the converse of this engine, namely one which applies work to extract heat from a cold reservoir and deliver heat to a warm reservoir. Such an engine is commonly known as a refrigerator.

Although this was derived for the Carnot cycle, it follows from the second law that this formula must be universal for reversible processes; for, if we would have a reversible process that would not satisfy the relation above, we could couple it to the Carnot cycle in a suitable way such as to fabricate a process that would violate the second law. For general processes, the law that we just derived takes the following form:

$$\oint \frac{dQ}{T} \; = \; 0 \text{ along reversible paths.}$$

This property now makes it possible to *define* the entropy, as a function of temperature and volume, by the following formula:

$$S(T_b, V_b) \; = \; S(T_a, V_a) + \int_a^b \frac{dQ}{T}$$

where the integral is taken along a reversible path. Indeed, it follows from the above that the answer does not depend on the choice of the reversible path. This is the macroscopic definition of the entropy. Note, by the way, that this definition only determines the entropy up to an additive constant.

As an example, let us compute the entropy of an ideal gas at a constant temperature T, so that the entropy will only be a function of the volume. Using the ideal gas law $PV = RT$, one gets

$$\begin{aligned} S(V_b) - S(V_a) \; &= \; \int_a^b \frac{dQ}{T} \; = \; \frac{1}{T}\Delta Q \; = \; \frac{1}{T}\int_a^b P\,dV \\ &= \; \frac{1}{T}\int_a^b \frac{RT}{V}\,dV \; = \; R(\log V_b - \log V_a). \end{aligned}$$

This result may already be related to the information-theoretic entropy, in the following way. A typical gas molecule may be found with equal probabilities in one of N_a 'cells'[12] in volume V_a, and in one of N_b cells in volume V_b. According to the definition of Shannon's entropy, the difference in entropies is proportional to $\log N_b - \log N_a = \log V_b - \log V_a$. Moreover, if the molecules are independent, we can just multiply by the number of molecules to get the difference of the total entropies. So at least in this simple case, the information-theoretic definition of entropy is in line with the thermodynamic one.

Very briefly, let us now describe the way the entropy is defined in statistical mechanics. The number of ways in which N_i indistinguishable particles may be divided over the g_i quantum states corresponding to a given energy level is given approximately (for $g_i >> N_i$) by $g_i^{N_i}/N_i!$. So the number of 'microstates' with N_1 particles on energy level E_1, N_2 particles on energy level E_2, etc., is given by

$$\Omega \; = \; \frac{g_1^{N_1}}{N_1!}\frac{g_2^{N_2}}{N_2!}\cdots.$$

[12] A cell is supposed to be just small enough to contain one molecule. Exactly how small this is fortunately turns out to be immaterial for the argument.

This is called the *thermodynamic probability*. Stirling's formula leads to the following approximate expression:

$$\log \Omega \;\approx\; \sum N_i \log \frac{g_i}{N_i} \;+\; \sum_i N_i.$$

The entropy is now defined as

$$S \;=\; k \log \Omega$$

where k is Boltzmann's constant. This quantity may also be interpreted as the logarithm of the volume of the part of the phase space that corresponds to a given macrostate. The increase of entropy appears in this framework as a statistical law, expressing the fact that a system when left to itself is most likely to go from states of low probability to states of high probability.

In an equilibrium state, one may assume the thermodynamic probability to be maximal under the constraints $\sum N_i = N$ (total number of particles) and $\sum N_i E_i = U$ (energy). Applying the standard Lagrange optimization method to the above approximate expression for $\log \Omega$ produces the distribution law

$$N_i \;=\; \alpha g_i e^{-\beta E_i}$$

where the constants α and β have to be determined from the side constraints. For an ideal gas, this leads to

$$\alpha \;=\; \frac{c_1}{VT^{3/2}}, \qquad \beta \;=\; \frac{c_2}{T}$$

where c_1 and c_2 are specific parameters for the gas. Computing $\log \Omega$, one obtains from this

$$\log \Omega \;=\; N \log V \;+\; \text{terms not depending on } V.$$

From the above, one sees that at least in the simplest case of the volume-dependence of the entropy of an ideal gas, this formula checks with the classical thermodynamic definition.

Entropy played a role in the birth of quantum mechanics; in fact Planck introduced his quantum postulate at the turn of the century as a device allowing him to obtain the correct formula for the entropy of a resonator. Of course, the device later turned out to be useful also for other purposes. In the modern axiomatization of the quantum theory, to which Von Neumann has contributed much, a 'normal state' is represented by a bounded linear operator ρ on a separable Hilbert space, satisfying trace $\rho = 1$. In terms of this axiomatization, our third definition of the physical entropy is

$$S_I(\rho) \;=\; -\text{trace}\, \rho \log \rho.$$

This is certainly similar to Shannon's definition of the entropy. It has been argued however, notably by LINDBLAD [10], that the above expression is invariant under the action of the Hamiltonians that describe motion in quantum mechanics, and so is not of use in finding the source of irreversibility. Lindblad proposes to use a different notion of entropy, which coincides with the above notion only for equilibrium states. He defines the *P-entropy* as

$$S(\rho; P) \;=\; \inf_{\gamma}\,[S_I(\rho(\gamma)) + \int \frac{dE}{T}(\rho;\gamma)]$$

where the infimum is taken over all paths that can be obtained by using a given set P of input vector fields, and that lead from the given state ρ to some equilibrium state $\rho(\gamma)$. This definition is remarkably close in spirit to the idea of storage functions that has been used in studies of dissipativity within system theory [19]. Also the idea of dependence on available vector fields is common in system theory, as noted by LINDBLAD [10, p. 14]. Which vector fields are available may depend on the current state of technology, and so the P-entropy as defined above is in principle a technology-dependent concept, although the state spaces in statistical mechanics are so large that the addition of a few more vector fields may not have much of an impact.

In conclusion, it is clear that there is a strong relation between the information-theoretic entropy and the entropy used in physics; in fact, for equilibrium states the two may be identified, up to the choice of 'unit of information'. Boltzmann's original picture is general enough to suggest also an information-theoretic interpretation of the physical entropy for non-equilibrium states, although the relation with Shannon's entropy (which is an *average* amount of information) is then less clear. In the discussion on reversibility and irreversibility, which inevitably accompanies the entropy concept in physics, contributions such as the one by Lindblad suggest that other elements besides information-theoretic ones may play a crucial role.

4. INFORMATION THEORY AND STATISTICS

The concept of information is basic in probability theory; it has even been argued that the concept of probability itself should be based on the notion of information, rather than vice versa [6]. In statistics, the word 'information' is used in various places, such as in the Fisher information matrix and in the information criteria used to determine the number of parameters in a model. The purpose of this section is to describe briefly some of the more conspicuous uses that have been made of entropy in statistics.

We begin with the so-called Kullback-Leibler distance between probability distributions, which may be introduced as follows (cf. [8]). Consider a stochastic variable with values in the discrete set $\{1, \cdots, n\}$. Suppose that we have the following hypotheses:

H_1: x is distributed according to $f_1 = (p_1, \cdots, p_n)$;

H_2: x is distributed according to $f_2 = (q_1, \cdots, q_n)$.

Let $P(H_j)$ denote our prior belief in H_j ($j = 1, 2$). How will these beliefs change on the basis of an outcome $x = i$? According to Bayes' rule, one has

$$P(H_j \mid x = i) \;=\; \frac{P(H_j)P(x = i \mid H_j)}{P(x = i)}.$$

This gives

$$\frac{P(H_1 \mid x = i)}{P(H_2 \mid x = i)} \;=\; \frac{P(H_1)}{P(H_2)}\frac{p_i}{q_i}$$

which expresses the (proportional) relation between the posterior beliefs in terms of the relation between the prior beliefs and the outcome i of the experiment. In order to describe the effects of successive experiments additively rather than multiplicatively, one can take logarithms and write

$$\log\frac{P(H_1 \mid x = i)}{P(H_2 \mid x = i)} \;=\; \log\frac{P(H_1)}{P(H_2)} + \log\frac{p_i}{q_i}.$$

The quantity $\log(p_i/q_i)$ can be interpreted as the evidence in favor of hypothesis H_1 as opposed to hypothesis H_2, due to outcome i. This quantity may of course be positive or negative. The *expected* evidence in favor of H_1 *in case* H_1 *is true* is now readily computed as

$$d_{\mathrm{KL}}(f_1, f_2) \;=\; \sum_{i=1}^{n} p_i \log\frac{p_i}{q_i}.$$

This is known as the 'relative entropy', the 'cross entropy', or the 'Kullback-Leibler distance' between the probability distributions f_1 and f_2. The term *distance* is used here in a rather wide sense: the Kullback-Leibler distance is not symmetric and does not satisfy the triangle inequality. It does have the most basic property of being nonnegative, though, as we should expect from a quantity that gives the expected evidence in favor of H_1 in case H_1 is true. A formal proof can readily be given[13]:

$$\sum p_i \log\frac{p_i}{q_i} = -\sum p_i \log\frac{q_i}{p_i} \geq -\log\sum p_i\frac{q_i}{p_i} = -\log\sum q_i = 0.$$

The inequality here uses the convexity of the logarithm, and the fact that the p_i's describe a convex combination. Since the logarithm is even strictly convex, the proof also shows that equality holds if and only if $p_i = q_i$ for all i.

As an example, we may consider the Kullback-Leibler distance between the joint distribution of two discrete variables x and y and the product of their marginal distributions. This distance is given by

$$\sum_i \sum_j P(x = i \wedge y = j) \log\frac{P(x = i \wedge y = j)}{P(x = i)P(y = j)}$$

[13] Note (for use later on) that the same inequality will hold if the q_i's sum up to less than one.

and may be interpreted as a measure of how quickly the dependence between x and y will show up in experiments consisting of drawing a value for x and one for y. The expression above is also exactly equal to the information in x about y as defined by Shannon.

The Kullback-Leibler distance can be related to Fisher information in the following way. Consider a set of parametrized densities $f(x, \theta)$, where the parameter vector θ takes values in $\mathbb{R}^k$, and define $d_{\mathrm{KL}}(\theta_1, \theta_2) = d_{\mathrm{KL}}(f(\cdot, \theta_1), f(\cdot, \theta_2))$. Choose a specific parameter value θ_0. The function $\theta \mapsto d_{\mathrm{KL}}(\theta_0, \theta)$ clearly has a minimum at $\theta = \theta_0$, and upon computing its Hessian one finds

$$\frac{\partial^2}{\partial \theta^2} d_{\mathrm{KL}}(\theta_0, \theta)\bigg|_{\theta=\theta_0} = G(\theta_0),$$

where $G(\theta_0)$ is the Fisher information matrix at θ_0. In other words, the Fisher information matrix will be close to being singular, so that even efficient estimators will have large variances, if the minimum of the Kullback-Leibler distance function at θ_0 is shallow. By letting the Fisher information matrix define a Riemannian metric on the parameter space, one may introduce a geodetic distance function between parameter values; this distance function is not the same as the Kullback-Leibler distance, but according to the formula above they are the same infinitesimally.

The inequality that expresses the nonnegativity of the Kullback-Leibler distance may also be written as

$$-\sum p_i \log q_i \geq -\sum p_i \log p_i. \tag{2}$$

As noted above, this inequality holds whenever $p_i \geq 0$, $q_i \geq 0$, $\sum p_i = 1$, $\sum q_i \leq 1$; moreover, equality holds if and only if $p_i = q_i$ for all i. The right hand side is, of course, the entropy of the probability distribution $(p_1, \cdots, p_n)$. The left hand side may be interpreted as the average number of bits per symbol in a code that uses $-\log q_i$ bits for symbol i. The requirement $\sum q_i \leq 1$ follows from the Kraft inequality for the lengths i of binary prefix codes[14]:

$$\sum 2^{-\ell_i} \leq 1$$

where $\ell_i = -\log q_i$ is the length of codeword i. The above inequality can therefore be interpreted by saying that the average code length is optimized by taking $q_i = p_i$. Turning this argument around, one may also say that optimizing the code for a given (long) sequence produces an estimate of the

[14] A *prefix code* is one in which no codeword is a prefix of another codeword, that is, no codeword can be extended with some symbols from the code alphabet to form another codeword. For binary codes, this may also be expressed as follows: if we associate with each codeword (say 010) the subinterval of $[0, 1]$ consisting of the numbers whose binary expansion begins with the given codeword (so all numbers of the form $.010\cdots$), then the resulting subintervals are nonoverlapping. This also proves the Kraft inequality, since the length of the subinterval associated with codeword i is $2^{-\ell_i}$, where ℓ_i is the length of the codeword, and the total length of the subintervals cannot exceed 1.

p_i's. This is one way of looking at the *minimum description length principle*, which has been suggested in particular by RISSANEN (see for instance [13]) as a rather universal strategy for estimation.

As was noted above, already in his original paper Shannon computed probability distributions that maximize the entropy under certain constraints. In a sense, the maximum-entropy distribution is the most acceptable one if only the data expressed by the constraints are given. Under certain reasonable-looking axioms, it can indeed be shown that *if* there is a general principle to determine probability distributions from 'testable information', then this should be the maximum-entropy principle [16]. The principle is popular in particular among Bayesians, because the maximization of the entropy gives a method to construct prior distributions; it does not hold a central place in mainstream statistics, however. An example of a maximum-entropy solution which has gained wide acceptance is Burg's construction of a stationary process when only the first k covariance matrices are given [2].

Maximum-entropy principles are used also outside statistics to find unique solutions (or, let us say, special solutions with interesting properties) for problems which otherwise would be underdetermined. One finds examples in interpolation theory [4], and in the H^∞-optimization which has become a major research direction in control theory [11].[15] These examples show that a concept of entropy may be defined and may lead to interesting results even in a deterministic context. What the *meaning* is of the entropy in such cases doesn't seem to have been discussed much.

5. ALGORITHMIC THEORY OF INFORMATION

The final topic in this introductory paper is the algorithmic theory of information, also known as the information-theoretic study of complexity. Basic references in this area include the papers by SOLOMONOFF [17], KOLMOGOROV [7], and CHAITIN [3], and the survey paper by ZVONKIN and LEVIN [21]. The approach hinges on the idea that the 'entropy' (or 'complexity') of a given sequence s can be determined as the length of the shortest program that computes s. At first sight, this idea looks unworkable, since we are all too familiar with the fact that the length of a program to compute some sequence (say the first 1000 digits in the decimal expansion of π) depends not only on that sequence, but also on the computer used. This problem can be overcome by using a 'universal' computer (actually a class of computers), and by being content with giving a definition of the entropy that contains an 'error term'.

In more detail, the algorithmic program proceeds as follows. First, one has to define what a computer is. The standard model is of course the Turing machine, as shown in Figure 4. The *program tape* is read-only, is scanned from left to right, and has finite length. The *work tape* is read/write, is infinitely long, and may be moved to the left and to the right. The computer is specified by giving, for each state and for each pair of squares being scanned: an action

[15] The appearance of 'minimum entropy' in the title of this reference is simply due to the fact that the authors reversed the usual sign convention.

FIGURE 4. Turing machine

(WRITE0, WRITE1, ERASE, LEFT, RIGHT, NEXT, STOP) and a next state. It is important to note that this specification may be given through a *finite* table. Before the beginning of a computation, there is a string present on the program tape which is called the *program*, and there may also be a string present on the work tape which is called the *input string*. A computation is *successful* if STOP occurs at the final square of the program tape. In that case, the *result* of the computation is the string that is left on the work tape from the read/write head to the first blank on the right. The result of a successful computation by computer C with program p and input string q is denoted by $C(p,q)$. The length of a string s is denoted by $|s|$. One can now define the class of computers that will be used in the definition of complexity.

> A computer U is an *optimal universal computer* if for every computer C there exists a constant k (depending on C) such that for all pairs of strings (p,q) the following holds: if $C(p,q)$ is defined, then there exists a p' with $|p'| \leq |p| + k$ and
>
> $$U(p',q) = C(p,q).$$

The crucial statement is: *there exists an optimal universal computer*. Indeed, Turing himself already showed how to construct a computer that is able to simulate any other computer when it is given a description of it (remember that a Turing machine is fully described by a finite table). If p_C is such a description and U is Turing's simulating computer, then we shall have $U(p_C p, q) = C(p,q)$ and so the condition of the theorem is satisfied if we take $p' = p_C p$. Obviously the length condition is fulfilled because $|p'| = |p| + |p_C|$.

Now, let U be a fixed universal computer. The *complexity* of a string s is defined as

$$H(s) \ = \ \min\{|p| \mid U(p,\varepsilon) = s\}$$

where ε denotes the empty string. In words, this is the length of the shortest program (running on U) that computes s when the input string is empty. The quantity $H(s)$ depends on the choice of a universal computer, but the difference between two values associated with two universal computers will be bounded by a constant that depends only on the computers and not on the strings.[16]

The next step is to define the *conditional complexity* of a string s *given* a string t. The definition that leads to the closest analogy with Shannon's information theory is given by CHAITIN [3]:

$$H(s \mid t) \;=\; \min\{|p| \mid U(p,t^*) = s\}$$

where

$$t^* \;=\; \arg\min\{|p| \mid U(p,\varepsilon) = t\}.$$

In other words, the conditional complexity of s given t is the length of a shortest program that computes s from t^*, where t^* itself is a shortest program to compute t.[17] This definition is chosen such that it becomes easy to construct a universal computer that will compute both s and t from a program t^*p, where t^* is a shortest program to compute t and p is a shortest program to compute s from t^*. Indeed, one can build a computer that will first simulate U, so that t is obtained, and then simulates U with t^* on the work tape and program p, in order to produce s. This construction proves that

$$H(s,t) \;\leq\; H(s) + H(t \mid s) + O(1).$$

With more work, one can even show that the inequality may be replaced by an equality:

$$H(s,t) \;=\; H(s) + H(t \mid s) + O(1).$$

This should be compared to formula (1) for Shannon's entropy. The analogy with concepts in Shannon's theory can be taken further by defining an 'information in s about t'

$$I(s:t) \;=\; H(t) - H(t \mid s),$$

and a 'probability of s'

$$P(s) \;=\; \sum_{U(p,\varepsilon)=s} 2^{-|p|}.$$

Then the expected relations hold up to an error term, such as

$$I(s:t) \;=\; I(t:s) + O(1)$$

and

$$H(s) \;=\; -\log P(s) + O(1);$$

[16] The fact that the complexity is only defined up to an error term limits its practical applicability. A more fundamental reason for the restriction of the complexity to the theoretical domain is that, as shown by Kolmogorov, the complexity is not effectively computable. One can work, however, with computable *approximations* of the complexity.

[17] The alternative is of course to let the complexity of s given t be the length of a shortest program that computes s from t itself. Besides being more obvious, this also has other advantages over Chaitin's definition.

the absence of an expectation operator here can be explained from the fact that we are considering the complexity of a single sequence.

The algorithmic theory assigns a complexity to single sequences, whereas Shannon's entropy is a statistical concept, defined for ensembles of sequences. The results that were just mentioned demonstrate a certain formal similarity, but one may also attempt to establish a connection in another way. To be specific, let us consider a stationary stochastic process $(x_1, x_2, \cdots)$ with values in $\{0, 1\}$, and fix a length n. The entropy $H(x_1, x_2, \cdots, x_n)$ is a *number*, whereas the complexity, which we now write $C(x_1, x_2, \cdots, x_n)$, is a *stochastic variable*. Taking the expectation of the complexity should bring us close to the entropy. And indeed, it was shown by LEUNG-YAN-CHEONG and COVER [9] that, under a computability condition on the marginal probability distributions, there is a constant k such that for all n

$$H(x_1, \cdots, x_n) \;\leq\; E\,C(x_1, \cdots, x_n) \;\leq\; H(x_1, \cdots, x_n) + k.$$

In particular, for the entropy rate of the process one has the relation

$$H \;=\; \lim_{n\to\infty} \frac{H(x_1, \cdots, x_n)}{n} \;=\; \lim_{n\to\infty} \frac{E\,C(x_1, \cdots, x_n)}{n}.$$

One can think of a shortest program for $(x_1, \cdots, x_n)$ as a code for the 'message' $(x_1, \cdots, x_n)$, and from this point of view it is not surprising that the expected value of the complexity is bounded below by the entropy, which gives on the average the shortest code length. It is much more surprising that a *universal* coding scheme can be built (i.e., one that is not given any information about the statistical properties of the source), which for a very large class of sources performs worse than the optimal one (which does use the statistical information) only by a constant that doesn't depend on the length of the message to be encoded.[18] Adaptive coding schemes are routinely used in data compression methods. A popular scheme is the one devised by ZIV and LEMPEL [20], which forms the basis of the UNIX command **compress**.[19]

REFERENCES

1. P. BRÉMAUD (1988). *An Introduction to Probabilistic Modeling.* Springer, New York.
2. J.P. BURG (1967). Maximum entropy spectral analysis. *Proc. 37th Meeting Soc. Exploration Geophysicists*, also in *Modern Spectrum Analysis*, D.G. Childers (ed.), IEEE Press, New York, 1978.
3. G.J. CHAITIN (1975). A theory of program size formally identical to information theory. *J. Association for Computing Machinery* **22**, 329–340.

[18] An analogy might be drawn with the field of adaptive control, in which one tries to construct controllers for dynamic systems to achieve certain goals such as closed-loop stability, with little information about the system to be controlled. The idea is to 'learn' the needed information along the way.

[19] Although the manual page doesn't mention it, I suspect that this is where the Z comes from in 'file.Z'.

4. H. Dym and I. Gohberg (1986). A maximum entropy principle for contractive interpolants. *J. Functional Analysis* **65**, 83–125.
5. J. Horgan (1992). Claude E. Shannon. *IEEE Spectrum* **29**, 72–75.
6. R.S. Ingarden and K. Urbanik (1962). Information without probability. *Colloquium Mathematicum* **9**, 131–150.
7. A.N. Kolmogorov (1965). Three approaches to the concept of the 'Amount of Information'. *Problems of Information Transmission* **1** (1), 1–7.
8. S. Kullback (1959). *Information Theory and Statistics.* Wiley, New York.
9. S.K. Leung-Yan-Cheong and T.M. Cover (1978). Some equivalences between Shannon entropy and Kolmogorov complexity. *IEEE Trans. Information Th.* IT-24, 331–338.
10. G. Lindblad (1983). *Non-Equilibrium Entropy and Irreversibility.* Reidel, Dordrecht.
11. D. Mustafa and K. Glover (1990). *Minimum Entropy H_∞ Control.* Lect. Notes Contr. Inform. Sci. **146**. Springer, Berlin.
12. A. Pais (1991). *Niels Bohr's Times, in Physics, Philosophy, and Polity.* Oxford University Press, New York.
13. J. Rissanen (1983). A universal prior for integers and estimation by minimum description length. *Ann. Stat.* **11**, 416–431.
14. C.E. Shannon (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423; 623–656.
15. C.E. Shannon and W. Weaver (1949). *The Mathematical Theory of Communication.* Univ. of Illinois Press, Urbana.
16. J. Skilling. Classic maximum entropy. In *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.). Kluwer Academic, Dordrecht, 1989, pp. 45–52, Proceedings of the 8th MaxEnt Workshop held at St. John's College, Cambridge, England, August 1–5, 1988.
17. R.J. Solomonoff (1964). A formal theory of inductive inference. *Information and Control* **7**, 1–22; 224–254.
18. N. Wiener (1948). *Cybernetics, or Control and Communication in the Animal and the Machine.* Technology Press MIT, Cambridge, Mass.
19. J.C. Willems (1972). Dissipative dynamical systems. Part I: General theory. *Arch. Rational Mech. Anal.* **45**, 321–351.
20. J. Ziv and A. Lempel (1978). Compression of individual sequences via variable-rate encoding. *IEEE Trans. Inform. Theory* IT-24, 530–536.
21. A.K. Zvonkin and L.A. Levin (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Math. Surveys* **25**, 83–124.