CHAPTER 4
ENTROPY AND INFORMATION

In the literature one finds the terms information theory and communication theory used interchangeably. As there seems to be no well-established convention for their use, here the work of Shannon, directed toward the transmission of messages, will be called communication theory and the later generalization of these ideas by Jaynes will be called information theory.

4.1 COMMUNICATION THEORY

Communication theory deals with the transmission of messages through communication channels where such topics as channel capacity and fidelity of transmission in the presence of noise must be considered. The first comprehensive exposition of this theory was given by Shannon and Weaver[1] in 1949. For Shannon and Weaver the term *information* is narrowly defined and is devoid of subjective qualities such as meaning. In their words:

> Information is a measure of one's freedom of choice when one selects a message. ... the amount of information is defined, in the simplest of cases, to be measured by the logarithm of the number of available choices.

Shannon and Weaver considered a message N places in length and defined p_i as the probability of the occurrence at any place of the i^{th} symbol from the set of n possible symbols and specified that the measure of information per place, H, should have the following characteristics

---

[1] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

1. The function H should be a continuous function of the probabilities, p_i's.
2. If all the p_i's are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n.
3. If A and B are two different independent events, and if we consider their joint occurrence, AB, as a compound event, then the uncertainty in AB is the sum of the separate uncertainties about A and B. That is, $H(AB) = H(A) + H(B)$.

Shannon and Weaver showed that only the following function satisfied these conditions

$$H = -K \sum p_i \log p_i \qquad \textbf{(4-1)}$$

Because K is an arbitrary constant, H is a relative measure and any convenient base can be chosen for the logarithm. If binary numbers are used $n = 2$ and $p_i = 1/2$ and the use of two-based logarithms yields

$$H = -K \left[\frac{1}{2} \log_2 \left(\frac{1}{2} \right) + \frac{1}{2} \log_2 \left(\frac{1}{2} \right) \right] = K$$

Setting K equal to unity we obtain $H = 1$, one unit of information per place — a bit. Alternatively, for a message in decimal numbers we have $n = 10$ and $p_i = 1/10$ and on using ten-base logarithms find

$$H = -10\,K \left[\frac{1}{10} \log \frac{1}{10} \right] = K$$

Again, setting K equal to unity will result in one unit of information per place.

The term H has been called the information per symbol, uncertainty per symbol, or the entropy. The terms information and uncertainty are not as incompatible as it would first appear. In fact, they are complementary in that the amount of information received is equal to the uncertainty removed on receipt of the message. For a message of N places the total amount of information contained is $N \cdot H$, the product of a quantity and an intensive factor.

Consider a message in the form of a three-digit decimal number (e.g., 329). It has just been shown that with ten-based logarithms $H = 1$ and therefore the amount of information in the message, $N \cdot H$, is 3 ten-base units. Now,consider this same number to be expressed in binary notation. Here it will be necessary to use a message length of ten places ($2^{10} = 1024$). At each place the probability is 1/2 and the use of ten-base logarithm gives

$$H = -K\left[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right] = K\log 2$$

Again, $K = 1$ and we obtain $H = 0.30103$ and $N \cdot H = 10(0.30103) = 3.01$ for a total of slightly over 3 ten-based units of information. If we had chosen to use two-based logarithms in Eq. (4-1), we would have obtained 9.966 and 10 two-based units[2] for the decimal and binary representations respectively. We would expect the amount of information in the message should be independent of the number system used and we see that this is essentially the case. The minor difference is due to a slight mismatch between the number systems: 1000 as compared to 1024. Because there are 1024 possibilities in the binary system and only 1000 in the decimal system,slightly more uncertainty is eliminated in specifying one member of a set of 1024 than one of a set of 1000.

Equation (4-1) was selected so that for any set of n symbols the maximum value of H occurs when all p_i are equal. This is the expected situation for a numerical message, but not for a message expressed in a language such as English. Because a language has structure, all symbols do not have the same frequency or probability of occurrence. Excluding punctuation, messages in English can be composed of the 26 characters of the alphabet and the space. If probabilities based on occurrence frequencies are used in Eq. (4-1), the value of H is 1.211 ten-based units. This is less than the value of 1.431 ten-base units computed from Eq. (4-1) using equal p_i of 1/27. The uncertainty per symbol is less in a structured language because we could make an educated guess.

---

[2] Note that $\log_2 M = \log M / \log 2$

In considering the transfer of information between two persons *A* and *B*, communication theory represents the chain as *A/a--b/B* where *a--b* is the physical part, the communication channel with transmitter *a* and receiver *b*. Communication theory is concerned only with the physical aspect of the exchange and takes no cognizance of meaning. At the receiver *b* it may be appropriate to consider the message as only one possibility out of a computable number of alternatives, but this certainly is not the attitude of *A* in originating the message.

Even in the simplest of cases, the quantity $N \cdot H$ does not give an unambiguous measure of information exchanged between *A* and *B*. For example, consider the various possibilities facing a present-day Paul Revere and his compatriot. By prior agreement the message could be coded to "0" or "1" (add one to get "one if by land, two if by sea"). If binary digits were used, the message would carry 1.0 two-base unit or 0.301 ten-base units of information. If the agreement were to send either "LAND" or "SEA" the message would carry 4.844 or 3.633 ten-base units respectively[3] and the amount of information would appear to depend on which route the British took. Would more information be imparted if the British were coming by land? This problem might be avoided if the prearranged code were "L" or "S". In this case either alternative imparts the same information, 1.211 ten-base units: but this is seen to be different from the information imparted from the use of binary digits. If no code were prearranged, the message might read "BRITISH COME BY LAND". This message of twenty places would convey 24.22 ten-base units of information. Note that in all cases the same amount of information was exchanged between *A* and *B* while the measurement in terms of the transmission from *a* to *b* varies from 0.301 to 24.22 units. Communication is a process and in communication theory it is restricted to the physical process connecting *a* and *b*. Communication theory

[3] We have seen that $H = 1.211$ ten-base units for messages in English and we obtain $4(1.211) = 4.844$ and $3(1.211) = 3.633$.

has been successfully applied and is a well-established tool of communication engineers.[4] On the other hand, information is a concept which is elusive and, as we have seen, is difficult, or perhaps impossible, to quantify.

The similarity between Shannon's H and the entropy of statistical mechanics, S_{SM}, can be seen by comparing Eq. (4-1) with Eq. (2-10)

$$S_{SM} = -k \sum P_i \ln P_i \qquad \textbf{(2-10)}$$

The k in Eq. (2-10) is not arbitrary, but is Boltzmann's constant and P_i is the probability that the system is in the i^{th} quantum state. Note that Eq. (2-10) was derived from a realistic model within an accepted conceptual structure and can be expected to be a reasonable reflection of reality. On the other hand, we have seen that Eq. (4-1) results from probabilistic considerations arising from the useful, but limited, concept of information. There is no physically realistic connection between H and S_{SM} and the designation of H as entropy can only lead to confusion. In quantum statistical mechanics it is possible to show that S_{SM} reduces to a heat effect divided by absolute temperature in conformance with the thermodynamic definition of entropy. However, no similar reduction is possible for H, in fact, the association of a heat effect with H would make no sense. Further, in examining the ramifications of the communication theory of Shannon and Weaver, one receives the impression that there is no distinct advantage in identifying H with entropy. None of the usual applications are exploited; in fact, the identification of system and surroundings would seem out of place. The association of H with entropy is regrettable for it has resulted in much confusion especially in regard to entropy and information.

The entropy as information interpretation has a long history dating back to Maxwell but the idea became more strongly established in 1929 with the work of Szilard. Here the concept of entropy of information was invoked by Szilard in

---

[4] See J.R. Pierce, *An Introduction to Information Theory*, 2nd ed., Dover Publications, New York, 1980.

order to save the second law of thermodynamics from the onslaught of a Maxwell demon of his devising. Szilard's work will be discussed in chapter 5.

4.2 JAYNES' FORMALISM

E.T.Jaynes has been the major force in generalizing certain aspects of Shannon's communication theory into what is sometimes known as information theory[5]. The cornerstone of this formulation is Jaynes' Principle which provides a means of determining the least-biased probability distribution. This principle merely states that, subject to the constraint that the sum of the P_i's is unity and other known constraints, the least-biased probability distribution is obtained by maximizing Shannon's H.

In applying Jaynes' formalism to quantum statistical mechanics, the entropy of statistical mechanics, S_{SM}, is used in place of Shannon's H.

$$S_{SM} = -k \sum P_i \ln P_i \qquad \textbf{(2-10)}$$

The usual constraints

$$\sum P_i = 1$$

and

$$\sum P_i E_i = U$$

are written and the method of Lagrangian multipliers is used to find the set of P_i that maximizes S_{SM} subject to these constraints. Once the probability distribution is found, it is used in

Eqs. (2-6) and (2-10) to obtain the internal energy and the entropy respectively. In this procedure the expression for the entropy is postulated and used to find probability distribution as opposed to the usual procedure of postulating the

---

[5] E.T. Jaynes, *Phys. Rev., 106*, 620 (1957) and *Phys. Rev., 108*, 171 (1957).

probability distribution and then determining the entropy. This inverted approach of the Jaynes formalism has been claimed to be more general than the approach of quantum statistical mechanics because it relies on logic rather than physics[6]. However, because the entropy in Jaynes' formalism is the entropy of physics, the claim is unconvincing.

Jaynes' Principle appears to possess considerable generality for statistical inference. It has proven useful in many fields and can be viewed as a tool which provides the best information available from a particular situation. However, it is unfortunate that this success has prompted a few enthusiasts[7] to propose the elevation of Jaynes' formalism to the status of a fundamental axiom of science and to state that it subsumes statistical mechanics. Such claims, together with giving the name entropy to Shannon's *H*, have served to establish the putative association of entropy with information. This association often leads to a subjective interpretation of entropy.

4.3 ENTROPY AND INFORMATION

The association of an increase in entropy with a decrease in information is an idea that dates back to the time of Maxwell[8] and the development of information theory (Jaynes' formalism) has seemed to reinforce this interpretation. According to this view, the condition $\Omega_2 > \Omega_1$,which leads by Eq. (2-11a) to an increase in entropy, represents an observer's loss of information about the microscopic state

---

[6] O. Costa de Beauregard and M. Tribus, *Helv. Phys. Acta, 47,* 238, 1974 also reprinted in H.S. Leff and A.F. Rex, *Maxwell's Demon: Entropy, Information, and Computing,* Princeton University Press, Princeton,N.J.,1990.

[7] See for example, M. Tribus, *Thermostatics and Thermodynamics,* D.Van Nostrand Co., Princeton, NJ, 1961.

[8] S. Brush, *The Kind of Motion We Call Heat,* North Holland Publishing Co., Amsterdam, 1976.

of the system. Accordingly, one reasons that there are more possibilities in state 2 and therefore the increase in Ω implies more uncertainty or a loss of information. This view presents two difficulties. First, because Ω is not a virtual observable quantity, it is doubtful that an observer could have access to this type of information. The information associated with Ω concerns not the system, but our description of the system. Second, it is unreasonable to believe that ΔS, a thermodynamic property change that depends on objectively determined macrostates, could also depend on microscopic information gained or lost by an observer.

In an effort to blunt the subjectivity criticism, Jaynes[9] has suggested the following carefully worded definition of information

> The entropy of a thermodynamic system is a measure of the degree of ignorance of a person whose sole knowledge about its microstate consists of the values of the macroscopic quantities X_i which define its thermodynamic state. This is a completely "objective" quantity, in the sense that it is a function only of the X_i, and does not depend on anybody's personality. There is then no reason why it cannot be measured in the laboratory.

Here, we will assume that the term knowledge is synonymous with information although in general usage they often have different connotations. Also, one wonders what type of knowledge would alleviate a person's ignorance of the microstate: virtual observable quantities such as positions and velocities or the identification of quantum states and their corresponding probabilities.

If Jaynes were speaking in quantum terms, there is a double dose of subjectivity here. First, we have introduced quantities such as Ω which are

---

[9] E.T. Jaynes, in *The Maximum Entropy Formalism*, R.D. Levine and M. Tribus, eds, M.I.T. Press, Cambridge, 1979.

mental constructs that relate to our description of the system rather than to the system itself. Second, we now say that the macroscopic behavior of the system, as reflected in the value of the entropy, is dependent on the extent of our knowledge of these model parameters. It would also be instructive to test the quantum interpretation through the use of Eq. (2-11a) that relates statistical entropy change to Ω_2/Ω_1 providing informational value could be assigned to the knowledge of Ω, a model parameter. This information should depend not on the value of Ω but on the range of numbers from which it is to be chosen. Thus, if Ω_1 and Ω_2 lie within the range 1 to N_{MAX}, they will have the same informational value regardless of their numerical values and there would be no change in our information concerning the system as it moves between macroscopic states. We reach the same conclusion by noting that the number of position coordinates and velocity components is always $6N$ regardless of the macroscopic state of the system — a constant amount of microscopic knowledge. Thus, from the viewpoint of "information" there would be no "entropy" change and we see that the concept of entropy as a measure of microscopic information is inconsistent as well as extremely subjective.