Entropy, information, and computation

J. Machta
Department of Physics and Astronomy, University of Massachusetts, Amherst, Massachusetts 01003-3720

(Received 20 May 1999; accepted 8 June 1999)

The relation between entropy, information, and randomness is discussed. Algorithmic information theory is introduced and used to provide a fundamental definition of entropy. The relation between algorithmic entropy and the usual Shannon–Gibbs entropy is discussed. © *1999 American Association of Physics Teachers.*

I. INTRODUCTION

In this article, I review the connections between entropy, information, and computation. The advent of mass-market computer technology means that students are now comfortable with the notion that information is physical and quantitatively measurable. Students are familiar with the idea that definite amounts of information may be stored in digital form on hard drives and other storage media and in dynamic memory. Thus, information can provide a useful handle for beginning statistical physics students struggling to understand the meaning of entropy. A discussion of the relationship between information and entropy also gives students an interdisciplinary perspective by showing that concepts central to statistical physics also appear in fields such as electrical engineering, computer science, and statistics.

The history of the relationship between entropy, information, and computation goes back to the first half of the 20th century with Szilard's analysis of Maxwell's Demon[1] and Shannon's work on communication theory.[2] Jaynes[3] and Brillouin[4] sought to place statistical mechanics on an information theoretic foundation. Shannon's definition of information is probabilistic and applies to ensembles of messages, just as the usual definition of entropy applies to statistical ensembles of microstates. A definition of the information content of individual objects was independently developed by Solomonoff,[5] Kolmogorov,[6] and Chaitin[7,8] and shown to be intimately related to Shannon's probabilistic definition. Based on this equivalence, Bennett[9] and Zurek[10,11] advanced the notion that the entropy of individual microstates of physical systems could be defined. This viewpoint is adopted in this article.

The foregoing developments are not usually treated in introductory statistical physics books. An exception is Baierlein's text,[13] which presents the subject of statistical mechanics at an elementary level using Shannon information theory as its basis. A recent resource letter in this publication[12] provides a bibliography on information theory in physics.

II. WHAT IS INFORMATION?

The information content, measured in bits, of a text document, audio recording, or data file is the number of ones and zeros needed to store the text, sound, or data using the most efficient digital encoding. As an example, consider a table of climate data. For simplicity, suppose that we have recorded only whether it has rained or not on a given day. A zero signifies ''no rain'' and a one signifies ''some rain.'' First, suppose we have a long data set for a rainy location like Seattle. We simplify the example by assuming that every day is independent of the previous days, and there is a 50% probability that it will rain on every day. The weather data we have recorded might as well have come from coin tosses and 8000 days of weather will require 8000 ones and zeros. A typical record might look like

0011101100101110001100110000111101101011110100000
    10010001101000…

We say that the information content of the data set is 8000 bits, 1 bit per day. In computer jargon 8 bits is 1 byte, so we would need 1 kilobyte of space on a hard disk to store the data. A crucial point is that because the data is random and without pattern, there is almost certainly no way to compress it to less than 1 kilobyte.

Next, consider the very different climate of Tucson, Arizona. Let us again suppose that every day is independent, but that it rains only 1 day out of 31 on the average. A typical record might look like

000000000000000000000000100000000000000000000000
    000000000000000…

Using the same encoding as before, we would need the same 1 kilobyte for 8000 days of weather. However, a typical record will be dominated by zeros, and there are more compact ways of storing the Tucson data by taking advantage of the knowledge that rain is rare. Here is one approach. Divide the record into 31-day intervals and for each interval indicate in binary on which days, if any, it rained. The binary code is as follows: the word 11111 marks the end of each 31 day interval, the word 00001 means rain on day one of the interval, 00010 means rain on day two, 00011 means rain on day 3, and so on, out to 11110 which means rain on day 31. An arbitrary weather record can be stored in this way. Because there will be 8000/31 month dividers and about 8000/31 rainy days, we will need about (8000/31)×5×2 bits or 0.32 bits per day. This amount is much better than 1 bit per day, although it is not the optimum compression. As we shall see, the best compression is about 0.21 bits/day.

To summarize, we say that the information content of a record is the number of bits (ones or zeros) needed to encode the record in the most efficient possible way. This definition is formalized by algorithmic information theory[8] and we will refer to information measured this way as algorithmic information content. An arbitrary sequence s of zeros and ones has an algorithmic information content $K(s)$ that is defined to be the size, in bits, of the smallest computer program that can be run to print out the sequence. The notion of program is used broadly here to include both the instructions for the computer and the data file. Thus, in the above weather example we should have also included the space taken by the instructions. For the example of Seattle, these instructions are very simple because the data were not compressed. For the Tucson example, the instructions involve the decompression of the data according to the stated rules. In both cases, the instructions are a negligible contribution for large data sets. Although there are ambiguities about the choice of computer used to print out the sequence, for a broad class of ''universal'' computers, these difference appear as additive constants and become unimportant for large data sets.

Algorithmic information is sometimes called algorithmic randomness. Some simple examples illustrate the relationship between information and randomness. First consider a string of N ones. This string has very little information content because the instructions to the computer are a simple do loop, ''For $i=1$ to N, print 1 and then stop.'' There is no additional data file to be read. A string of N ones is completely ordered and not at all random. On the other hand, the result of N coin tosses has an algorithmic information content that is typically about N because there is usually no compression of the data possible, and it is necessary to store the entire data file. For a typical random string the instructions are simple, ''For $i=1$ to N, print record i in the data file and then stop,'' but the data file has a size of order N. The results of random processes usually have high information content. It is possible, however, for a coin to be tossed N times and yield N heads or some other ordered pattern with little information content. Not all sequences that appear to be random have a high algorithmic information content. For example, the first N bits of the binary expressions for π or e look random and pass most statistical tests of randomness, but they have little algorithmic information content because there are concise algorithms for computing these numbers to arbitrary precision. On the other hand, any sequence that has an algorithmic information content comparable to the length of the sequence will appear to be random and will pass all statistical tests of randomness. Thus, algorithmic information is sometimes called algorithmic randomness. Randomness and information are formally the same thing. If we want to emphasize the utility or value of some data, we speak of information content. If we want to emphasize a lack of pattern or order in some data, we speak of randomness. A high algorithmic information content does not imply that the data are meaningful or useful.

There is a second definition of information that is formally the same as the standard definition of entropy in statistical mechanics. This definition is due to Shannon[2] and arose in his analysis of the capacity of communications channels. Suppose that there are W possible messages labeled s_i, $i=0,1, \quad W-1$, that can be sent and that the probability that s_i is sent is p_i. Then, the information content I per message transmitted is

$$I=-\sum_{i=0}^{W-1} p_i \log_2 p_i . \tag{1}$$

We call I in Eq. (1) the Shannon information.

The Shannon information has an additive property. If a message is composed of pieces that are statistically independent, then the information content of the entire message is the sum of the information content of the pieces. For the example of weather data, each day's record is statistically independent, so the information content of the entire record is simply 8000 times the information content from 1 day. For 1 day there are two possible messages, 0 for no rain and 1 for some rain, with probabilities p_0 and p_1. For Seattle, $p_0=p_1=1/2$, so according to the Shannon formula, $I=-[(1/2)\log_2(1/2)+(1/2)\log_2(1/2)]=1$ bit per day. The Shannon information and the algorithmic information agree at 1 bit per day. For the Tucson example, $p_0=30/31$ and $p_1=1/31$, so $I=-[(30/31)\log_2(30/31)+(1/31)\log_2(1/31)]=0.21$ bits per day.

On the face of it, the two definitions of information appear very different. The Shannon information is calculated from a definite formula involving probabilities. It is not applicable to a single sequence of ones and zeros but only to a statistical ensemble of such sequences. Algorithmic information content is applicable to a single sequence, but cannot be calculated by a formula or any definite procedure because it depends on finding the best way to compress the data.[14] Nonetheless, there is a close connection between the two definitions. Given an ensemble of possible messages, s_i with associated probabilities p_i, a fundamental result of algorithmic information theory is that

$$\langle K\rangle \cong I, \tag{2}$$

where the average algorithmic information is defined by

$$\langle K\rangle=\sum_i p_i K(s_i), \tag{3}$$

and I is the Shannon information defined in Eq. (1). Equation (2) is an approximate equality. Differences between the right and left sides come from the choice of universal computer and from a term involving the algorithmic information required to specify the probabilities.[15] For equilibrium statistical mechanics, the probabilities are concisely defined and typical amounts of information are large, and hence Eq. (2) is essentially exact.

III. WHAT IS ENTROPY?

In thermodynamics, entropy is an extensive quantity associated with a system in equilibrium. Entropy may be added or removed from a system by adding or removing heat. If the system remains near equilibrium, the entropy change is equal to the heat transfer divided by the absolute temperature,

$$\Delta S=Q/T, \tag{4}$$

where S is the entropy, Q is the heat transfer into the system, and T is the absolute temperature. For irreversible processes, the entropy must obey the second law which says that the entropy of an isolated system may never decrease.

The conventional ensemble definition of entropy in statistical mechanics is due to Gibbs. Suppose a system can be in one of a large number of microstates. In quantum mechanics, a microstate is specified by giving a complete list of the quantum numbers of the system. In classical mechanics, it is the location in phase space— the positions and momenta of all the particles in the system. Suppose that the probability of a system being in microstate i is p_i. In the canonical ensemble, this probability is

$$p_i=\frac{e^{-E_i/k_BT}}{Z}, \tag{5}$$

where E_i is the energy of the microstate, k_B is Boltzmann's constant, and Z is the partition function, which is needed to normalize the probabilities. In the microcanonical ensemble, all the p_i are equal for microstates with E_i in a narrow range near the thermodynamic energy while outside of that range, the p_i vanish.

The Gibbs entropy is given by

$$S=-k_B\sum_i p_i \ln p_i, \tag{6}$$

where the summation is over all possible microstates available to the system. Note that this definition encompasses the earlier definition proposed by Boltzmann for the microcanonical ensemble,

$$S=k_B \ln \Omega, \tag{7}$$

where Ω is the ''statistical weight,'' the number of microstates having energies within a narrow range near the thermodynamic energy. Because in the microcanonical ensemble, each state in the energy range has the same probability, this probability must be $1/\Omega$ and because there are Ω equal terms in the sum, Eq. (6) reduces to Eq. (7). Standard arguments show that the Gibbs entropy has the properties required of entropy by thermodynamics.

Extensive quantities such as the number of particles or the energy have definite values for individual microstates, but the Gibbs entropy is defined only for statistical ensembles. Indeed, if a macroscopic system could be prepared in a definite microstate, the unsettling implication of Eq. (6) is that it would have zero entropy. Is there a way to define the entropy of an individual microstate of a system?

IV. ENTROPY AND INFORMATION

A comparison of Eqs. (1) and (6) reveal that the Shannon information and the Gibbs entropy are formally the same except for a constant factor $\kappa=k_B/\log_2 e=9.57\times 10^{-24}$ $\mathrm{J/K^{-1}\,bit^{-1}}$. How should we interpret this coincidence?

Brillouin and Jaynes developed the point of view that entropy is a measure of our lack of information about the microstate of a system. Probabilities must be assigned to microstates because we do not know what microstate the system is in. The missing information is the information that would be gained if a complete measurement is made on the system so that the exact microstate is known. The information gained in this way is, on average, the Shannon information or, up to a constant, the Gibbs entropy. The correct assignment of probabilities should be made in such a way that no unjustified assumptions about the system are built into the probabilities. Probabilities are assigned by building in what is known about the system and then maximizing the missing information. This prescription for assigning probabilities is useful in various applications of statistics and has become known as the maximum entropy principle. As applied to statistical mechanics, it yields the microcanonical or canonical ensembles. For example, if the Gibbs entropy is maximized holding the average energy fixed, the resulting distribution is the canonical ensemble.

The thesis that entropy is missing information is unsatisfactory because it makes entropy a subjective rather than an objective property of physical systems. I favor a viewpoint espoused by Bennett and Zurek that makes entropy an objective property of physical systems. Suppose that a complete description of the microstate s of a system has an algorithmic information content $K(s)$. Define the algorithmic entropy of microstate s as $\kappa K(s)$. Equation (2) insures that the ensemble-averaged algorithmic entropy will be the same as the Gibbs entropy and thus a faithful representation of the thermodynamic entropy.

The algorithmic entropy is now taken as the fundamental theoretical definition of entropy. For calculations we will still use ensemble methods but now with algorithmic entropy as a foundation. Because we have no information about a system other than a few thermodynamic variables, we choose a probability distribution according to the maximum entropy principle. Given these probabilities we calculate averages of physical quantities. The average entropy, defined in Eq. (3), is evaluated using Eq. (2). This way of thinking puts entropy on nearly the same footing as other extensive quantities such as energy. The entropy has a definite (though uncomputable) value for a physical system but, because of our lack of information, we actually calculate an average value over an ensemble. Because statistical mechanics ensembles for macroscopic systems are very sharply peaked, the average is a very accurate estimate of the actual value.

To illustrate these ideas, consider the entropy of an ideal monatomic quantum gas obeying Maxwell–Boltzmann statistics. The gas consists of N atoms of mass m in a box of volume V. The microstates are defined by the occupancies of single particle quantum levels in the box. The Sakur–Tetrode formula for the entropy is

$$S=k_B N \ln\left[\frac{V}{N}\left(\frac{mk_BT}{2\pi\hbar^2}\right)^{3/2}\right]+\frac{5}{2}k_B N. \tag{8}$$

For 1 mole of ^{4}He at 300 K confined to one liter, the entropy is about 100 J/K. The interpretation of this result is that a complete description of a single microstate would, on average, require $S/\kappa=100/9.57\times 10^{-24}\approx 10^{25}$ bits, or about 17 bits per atom. Note that the natural microscopic unit for entropy is the bit. The Sakur–Tetrode formula is usually derived from the Gibbs entropy but can also be derived directly from the algorithmic entropy.[10]

The algorithmic view is useful in clarifying situations where some of the degrees of freedom of a physical system are ''information bearing.'' As a specific example, consider the analysis of a 1-gigabyte hard disk drive. The heart of this device is a metal disk coated with a film of magnetic material. Like any macroscopic object, this disk has a huge number of degrees of freedom. Of these degrees of freedom, a tiny fraction, roughly 8×10^9 (corresponding to 1 gigabyte), are information-bearing degrees of freedom which can be read or modified. The information-bearing degrees of freedom are collective variables, referring to the magnetization of many electrons in a specific region on the disk. Reading (writing) is done by a head that rides over the surface of the disk measuring (changing) the magnetization. The information-bearing degrees of freedom contribute to the algorithmic entropy in exactly the same way as all the other degrees of freedom.

Suppose that the hard drive is initially filled with a record which is the result of 8 billion coin tosses. The entropy associated with the information-bearing degrees of freedom will be $\kappa(8\times 10^9)$ bits. Suppose that the disk is erased, meaning that the disk is restored to some simple state with very little algorithmic information. We conclude that the entropy of the drive has decreased. To satisfy the second law, an equal or greater increase in entropy must have occurred elsewhere. If the process occurs near equilibrium at temperature T, then according to Eq. (4) a tiny amount of heat

$\kappa(8\times 10^9\text{ bits})(300\text{ K})=2.3\times 10^{-11}$ J must be released. This release of heat requires an expenditure of the same amount of free energy. In practice, much more free energy than this amount is dissipated when information on a hard drive is erased because many other dissipative processes occur. However, as a matter of principle, the analysis shows very generally that there is a minimum dissipation of κT whenever a bit of information is erased in an environment at temperature T. This result is known as *Landauer's principle*.[16,17] Other aspects of information processing can, at least in principle, be carried out reversibly. Reading, copying, and computing can all be carried out without dissipation,[9] although in practice, each of these processes dissipates much more than κT per elementary step.

Having made the argument that information and entropy are fundamentally equivalent, it is useful to distinguish between degrees of freedom that are under our control and easily measured and degrees of freedom that are not under our control and not easily measured. Although this distinction is fuzzy and changes as technology advances, it is nonetheless useful to associate the term information with the controlled degrees of freedom and the term entropy with uncontrolled degrees of freedom. Erasing information is a process in which information/entropy is moved from controlled to uncontrolled degrees of freedom. Thus, by definition, erasure is an irreversible process.

The algorithmic approach to entropy does not resolve the fundamental questions surrounding the second law and the arrow of time. Algorithmic entropy/information is essentially conserved by classical or quantum dynamics because of time-reversal invariance. Small perturbations from outside the system or other sources of decoherence are required to explain the increase in information/entropy during the equilibration of an isolated system.

V. SUMMARY

We have seen that the notions of entropy, information, and randomness are equivalent and can be defined for individual microstates of physical systems using the ideas of algorithmic information theory. Algorithmic information content is the number of bits required to store a record in the most compressed possible form. The algorithmic definition of entropy is equivalent to the Gibbs ensemble definition. The ensemble approach is required for most calculations but the algorithmic viewpoint has some conceptual advantages. The algorithmic approach gives entropy an objective meaning, and it clarifies the analysis of systems with information handling abilities. Landauer's principle applies to such systems and states that κT free energy must be dissipated when one bit of information is erased in an environment at temperature T.

ACKNOWLEDGMENTS

I thank Wojciech Zurek for carefully reading the manuscript and suggesting improvements. I also thank Harvey Gould and Jan Tobochnik for carefully editing the manuscript and for creating this special issue of the American Journal of Physics.

[1] L. Szilard, ''On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings,'' Z. Phys. **53**, 840–856 (1929). Translation by A. Rapoport and M. Knoller, reprinted in *Quantum Theory and Measurement*, edited by J. A. Wheeler and W. H. Zurek (Princeton U. P., Princeton, 1983).

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

[3] E. T. Jaynes, ''Information theory and statistical mechanics,'' Phys. Rev. **106**, 620–630 (1957).

[4] L. Brillouin, *Science and Information Theory* (Academic, New York, 1956).

[5] R. J. Solomonoff, ''A formal theory of inductive inference, Parts I and II,'' Inf. Control. **7**, 1–22; **7** 224–254 (1964).

[6] A. N. Kolmogorov, ''Three approaches to the quantitative definition of information,'' Probl. Peredachi Inf. **1** (1), 1–7 (1965). English translation: Probl. Inf. Transm. **1**, 1–7 (1965).

[7] G. J. Chaitin, ''On the length of programs for computing finite binary sequences,'' J. Assoc. Comput. Mach. **13**, 547–569 (1966).

[8] G. J. Chaitin, ''Randomness and mathematical proof,'' Sci. Am. **232**, No. 5, 46–52 (1975), available at http://www.cs.auckland.ac.nz/CDMTCS/chaitin/sciamer.html.

[9] C. H. Bennett, ''The thermodynamics of computation—a review,'' Int. J. Theor. Phys. **21**, 905–940 (1982).

[10] W. H. Zurek, ''Algorithmic randomness and physical entropy,'' Phys. Rev. A **40**, 4731–4751 (1989).

[11] W. H. Zurek, ''Algorithmic information content, Church–Turing thesis, physical entropy, and Maxwell's Demon,'' in *Complexity, Entropy and the Physics of Information*, Santa Fe Institute Studies in the Sciences of Complexity, Vol. VIII, edited by W. H. Zurek (Addison-Wesley, Redwood, 1990), pp. 73–89.

[12] W. T. Grandy, ''Resource letter ITP-1: Information theory in physics,'' Am. J. Phys. **65**, 466–476 (1997).

[13] R. Baierlein, *Atoms and Information Theory* (Freeman, San Francisco, 1971).

[14] A fundamental feature of algorithmic information is that there can be no algorithm that can be used to calculate the algorithmic information content of an arbitrary string of ones and zeros. This result is closely related to Gödel's incompleteness theorem; see Ref. 8.

[15] A probability distribution is concisely defined if there is a short program to compute the probability of any string. Trivial probability distributions may not be concisely defined. For example, the distribution, $p(\tilde{s})=1$ for some particular string $\tilde{s}$ and zero otherwise, is not concisely defined if $\tilde{s}$ has a high information content. The canonical ensemble of statistical mechanics is concisely defined by Eq. (5) and a procedure for calculating the energy.

[16] R. Landauer, ''Irreversibility and heat generation in the computing process,'' IBM J. Res. Dev. **5**, 183–191 (1961).

[17] R. Landauer, ''Fundamental physical limitations of the computational process,'' in *Computer Culture: The Scientific, Intellectual, and Social Impact of the Computer*, edited by H. Pagels, Ann. (N.Y.) Acad. Sci. **426**, 161–170 (1985).