

Some Observations on the Concepts of Information-Theoretic Entropy and Randomness

Jonathan D.H. Smith

Department of Mathematics

Iowa State University

Ames, IA 50011, USA

E-mail: jdsmith@math.iastate.edu

URL: <http://www.math.iastate.edu/jdsmith/>

Received: 15 February 2000 / Accepted: 11 January 2001 / Published: 1 February 2001

Abstract: Certain aspects of the history, derivation, and physical application of the information-theoretic entropy concept are discussed. Pre-dating Shannon, the concept is traced back to Pauli. A derivation from first principles is given, without use of approximations. The concept depends on the underlying degree of randomness. In physical applications, this translates to dependence on the experimental apparatus available. An example illustrates how this dependence affects Prigogine's proposal for the use of the Second Law of Thermodynamics as a selection principle for the breaking of time symmetry. The dependence also serves to yield a resolution of the so-called "Gibbs Paradox." Extension of the concept from the discrete to the continuous case is discussed. The usual extension is shown to be dimensionally incorrect. Correction introduces a reference density, leading to the concept of Kullback entropy. Practical relativistic considerations suggest a possible proper reference density.

Keywords: information-theoretic entropy; Shannon entropy; Martin-Loef randomness; self-delimiting algorithmic complexity; thermodynamic entropy; Second Law of Thermodynamics; selection principle; wave equation; Gibbs Paradox; dimensional analysis; Kullback entropy; cross-entropy; reference density; improper prior.

1 Introduction

The aim of this note is to discuss certain aspects of the concepts of information-theoretic entropy and randomness, particularly concerning their use in physics. Many of the points to be raised are “well-known” in certain communities, but not necessarily appreciated outside those communities. Thus the note is intended to further the dialogue about this deep and often controversial subject.

For a system with a finite set $\xi = \{C_1, \dots, C_r\}$ of macrostates, the (*information-theoretic*) entropy is defined to be

$$H(\xi) = - \sum_{i=1}^r p_i \log p_i, \quad (1)$$

where p_i is the probability of the i -th state C_i , and where $p_i \log p_i$ is zero if p_i is zero. The quantity (1) is often called “Shannon entropy”, with reference to Shannon’s (1948) work. However, Tolman (1938) attributes the definition to Pauli (1933), where one already finds its application to statistical mechanics and relation to thermodynamic entropy, as outlined briefly in Section 3 below. (This pre-history spoils the otherwise amusing anecdote that use of the word “entropy” for (1) is the result of a joke told by Neumann to Shannon.) Many derivations of (1) assume a large phase space, and use Stirling’s formula to derive (1) and related formulae as the limit of combinatorial counts involving factorials [cf. Rumer and Ryvkin (1977)]. In Section 2, a derivation of (1) from first principles is given, inspired by Martin-Löf’s (1966) definition of the randomness of a sequence. We interpret the sequence as the sequence of outcomes of an experiment. In this approach, the concepts of entropy, probability, and randomness turn out to be mutually equivalent, in the sense that a conceptual foundation for any one yields a foundation for the other two. The derivation of (1) is completely valid for all discrete phase spaces, even small ones. No approximations are involved.

An important and easily overlooked aspect of randomness is its dependence on the experimental setup, and on the sophistication of the apparatus involved. In modern refinements of Martin-Löf’s definition of randomness, this dependence is modelled by the complexity class of the algorithms available to detect patterns in sequences. A macroscopic physical counterpart of this dependence is discussed in Section 4, considering the applicability of the Second Law of Thermodynamics as a “selection principle” in the sense of Prigogine and George (1983). Similar considerations of the power of the available apparatus serve to resolve the so-called “Gibbs Paradox”. One may summarize by saying that thermodynamic entropy measures the accessibility of energy to extraction by the machinery involved. More sophisticated machinery may render more energy accessible to extraction, and will thus correspond to a different count of thermodynamic entropy from that obtained when using less sophisticated apparatus.

The final topic discussed is the problem of extending the formula (1) from a finite sum to a continuous integral for the purpose of physical applications. It is not sufficient just to regard sums such as (1) as being Riemann sums or other similar approximations of integrals. Dimensional analysis forces one to replace (1) by a relative entropy or “cross-entropy”. The choice of reference distribution then depends on the experimental setup, in particular on the time available for the experiment.

2 Randomness, probability, and entropy

Much of the controversy surrounding the information-theoretic entropy concept has arisen from the long-standing lack of a solid basis for the understanding of randomness. Measure theory provided a sound axiomatic foundation for probability theory from the pure mathematics standpoint, but begged the question of the applicability of the axioms. Although the issue has by no means been completely resolved, and much remains to be done (particularly in the area of quantum probability and randomness), the general approach of Martin-Löf (1966) (see also Li and Vitanyi, 1997) suggests a practicable definition for the classical case along the following lines.

Consider a scientific experiment which may have any one of N possible outcomes. By convention of modern science, a “scientific” experiment has to be repeatable and reproducible. The experiment is said to be *random* if no statistical test available to the experimenter can detect any pattern in repeated outcomes. It is very important to note that this definition is contingent on the power of the apparatus available. Rolling a die under casino conditions should be random. On the other hand, if equipment such as a high speed precision camera were available, it would be possible to predict the outcome of each roll from the initial motion of the die, and the experiment would no longer be random. A subtler distinction arises when rolling a very slightly loaded die under casino conditions. If the number of rolls required for even a sophisticated statistical test to reveal the bias is in excess of the number of rolls sufficient to wear the spots off the die, then the experiment would still qualify as random.

The basic concepts of entropy and probability follow from the concept of randomness. For a random experiment with N outcomes, the (*natural*) *entropy* is

$$H = \log N \quad (2)$$

(using natural logarithms) or the (*binary*) *entropy*

$$H = \log_2 N \text{ bits} \quad (3)$$

using logarithms to base 2 and bits as units. (Occasionally logarithms to base 10 are used, with Hartleys as units.) The *probability* $\pi(x)$ of any one particular outcome x of the random experiment

is

$$\pi(x) = N^{-1}. \quad (4)$$

The probability $\pi(x)$ and natural entropy H are connected by the mutually inverse relationships

$$(a) \quad H = -\log \pi(x), \quad (b) \quad \pi(x) = \exp(-H). \quad (5)$$

Using binary entropy H , these take the form

$$(a) \quad H = -\log_2 \pi(x), \quad (b) \quad \pi(x) = 2^{-H}. \quad (6)$$

The probability $\pi(x)$ represents a fair stake to buy into the following game: win one unit if the outcome of the experiment is x . Randomness of the experiment means that there are no winning strategies in this game. The entropies (2) and (3) measure one's ignorance about the outcome of the experiment. If you let someone else run the experiment, and instead question them afterwards as to what the outcome was, then the number of yes/no questions required to elicit the outcome would be given by (3).

The above model is too narrow for general use, when one wishes to deal with non-random experiments. (By this, we mean general experiments that traditional statisticians might call "random," but having a non-uniform finite rational probability distribution.) These may be modelled using an underlying random experiment (as in the preceding paragraph) whose set of outcomes, called the *phase space*, has N elements. The phase space is completely partitioned into a set $\xi = \{C_1, \dots, C_r\}$ of mutually exclusive subsets called *states*. The partition ξ represents the non-random experiment (also denoted ξ) of sampling an outcome x from the phase space and locating the state C_i in which it lies. If the state C_i contains n_i outcomes of the underlying random experiment, each of whose outcomes has probability N^{-1} according to (4), then the *probability* $p(C_i)$ of the state C_i is given as

$$p(C_i) = n_i N^{-1}. \quad (7)$$

The state C_i may be regarded as a random experiment in its own right: select an outcome from C_i . The entropy of this random experiment, according to (2), is

$$H(C_i) = \log n_i. \quad (8)$$

If you perform experiment ξ and obtain the result C_i , then your ignorance will have been reduced by $\log N - \log n_i = -\log p(C_i)$. This happens with probability $p(C_i)$. Thus the average loss of ignorance or gain in knowledge obtained on performing experiment ξ is its *entropy*

$$H(\xi) = -\sum_{i=1}^r p(C_i) \log p(C_i). \quad (9)$$

(Of course, one may take logarithms to base 2 and quote $H(\xi)$ in bits.) The mathematical discipline of measure theory extends the definitions of probability and entropy to appropriate infinite phase spaces, where “counting outcomes” may be replaced by “measuring volumes”. The entropy $H(\xi)$ satisfies the inequality

$$0 \leq H(\xi) \leq \log r. \quad (10)$$

Equality obtains on the left in (10) if and only if $p(C_i) = 1$ for some i : if you already know in advance that ξ will come up with state C_i , then you gain no knowledge by performing the experiment. Equality obtains on the right in (10) if and only if $p(C_i) = r^{-1}$ for each i : the most informative experiments are those designed so that all their different outcomes are equally likely. In particular, randomness of an experiment is characterized by its entropy. Moreover, the three concepts of “entropy”, “probability”, and “randomness” turn out to be equivalent, in the strict mathematical sense that establishment of any one leads to establishment of the other two. For example, formulae such as (5) establish mutual connections between entropy and probability. In complexity theory, it is an entropy measure, namely the (relativised) self-delimiting algorithmic complexity, which is usually taken as basic. Probability is then obtained via (6) (Uspenskii, Semenov, and Shen', 1990). The complexity class of the algorithms invoked corresponds to the power of the apparatus used in the statistical tests for randomness above.

3 Statistical mechanics: the canonical ensemble

For the experiment $\xi = \{C_1, \dots, C_r\}$ considered in the previous section, absolute randomness – complete ignorance about the outcome – was characterized by unconstrained maximization of the entropy $H(\xi)$, attaining the value $\log r$ according to (10). In practice it may be possible to assign a numerical value E_i to each state C_i , e.g. the number of spots on the face of a die or an energy in electron volts. If the expected value

$$E = \sum_{i=1}^r p(C_i) E_i \quad (11)$$

is known, then the non-negative probabilities $p(C_i)$, which have to satisfy the relationship

$$1 = \sum_{i=1}^r p(C_i), \quad (12)$$

are determined by the assumption of relative randomness: no pattern is discernible in repeated outcomes except for maintenance of the fixed value E . This is equivalent to maximization of the entropy $H(\xi)$ subject to the constraints (11) and (12). Setting $g_1 = E - \sum_{i=1}^r p(C_i) E_i$, $g_2 =$

$1 - \sum_{i=1}^r p(C_i)$, and $f = H(\xi) + \beta g_1 + \lambda g_2$ with Lagrange multipliers β, λ , the stationarity conditions $\partial f / \partial p(C_i) = 0$ lead to $\log p(C_i) = -\beta E_i - (1 + \lambda)$ or $p(C_i) = \exp(-\beta E_i) / \exp(1 + \lambda)$. Substituting into (12), noting that λ is independent of i , one obtains

$$p(C_i) = Z(\beta)^{-1} \exp(-\beta E_i) \quad (13)$$

with the *partition function* or *Zustandsumme*

$$Z(\beta) = \sum_{i=1}^r \exp(-\beta E_i). \quad (14)$$

The fixed value E from (11) is recovered as

$$E = -\frac{d \log Z(\beta)}{d\beta}; \quad (15)$$

on the other hand this equation may yield β if the partition function is continuous and strictly logarithmically convex. The entropy (9) may be recovered via

$$-\sum_{i=1}^r p(C_i) \log p(C_i) = \sum_{i=1}^r p(c_i) [\beta E_i + (1 + \lambda)] \quad (16)$$

as

$$H(\xi) = \beta E + \log Z(\beta). \quad (17)$$

Note that (13)–(17) are just consequences of the assumption of randomness of ξ relative to (11) and the attribution of the numerical value E_i to each state C_i . There is no inherent reason for such an experiment ξ to be inappropriate as a model in certain "non-equilibrium" circumstances. (In many cases additional numbers F_i, G_i , etc. may be assigned to each state, with known expected values F, G , etc. The Lagrange multiplier method readily extends to such cases, using vectors in place of scalars.) An experiment ξ with random probabilities subject to (11) is called a *canonical ensemble*, since it generalizes the models of that name in (equilibrium) statistical mechanics. A correspondence with thermodynamics arises when the E_i and E are energies in suitable units. Introducing *Boltzmann's constant* k , the *thermodynamic entropy* is

$$S = kH. \quad (18)$$

The *temperature* is

$$T = 1/k\beta. \quad (19)$$

The *thermodynamic potential* is

$$\Psi = \log Z(\beta). \quad (20)$$

The *Helmholtz free energy* is

$$F = -kT\Psi. \quad (21)$$

Equation (17) then reduces to

$$F = E - TS. \quad (22)$$

Note that (13) and (19–21) yield the probability of the state C_i as

$$p(C_i) = \frac{e^{-E_i/kT}}{e^{-F/kT}}. \quad (23)$$

This gives an immediate derivation of the formulae of kinetic theory such as those in Chapter 42 of the “red book” (Feynman, Leighton, and Sands, 1963).

4 Dependence on the apparatus

In Section 2, the mutual equivalence of the concepts of entropy, probability, and randomness was adumbrated. It was also pointed out that these concepts are contingent upon the power of the apparatus available. Prigogine and George (1983) have suggested the use of the Second Law of Thermodynamics as a *selection principle*, breaking the microscopic time symmetry of physical laws expressed by differential equations. Suppose that a differential equation is invariant under time reversal, say because it only involves even time derivatives. Then it may formally possess solutions going both forward and backward in time. Somehow, the solutions going backward in time are to be eliminated. The idea of the Second Law as a selection principle is to dismiss the backward solutions as being “improbable”. Echoing the point made early in Section 2, we wish to emphasize here that the improbability of the backward solutions also depends critically on the sophistication of the experimental apparatus.

The use of the Second Law of Thermodynamics as a selection principle may be illustrated by the wave equation, as the differential equation

$$\frac{\partial^2 \phi}{\partial r^2} + \frac{2}{r} \frac{\partial \phi}{\partial r} = \frac{1}{c} \frac{\partial^2 \phi}{\partial t^2} \quad (24)$$

for the spherically symmetric air pressure $\phi(r, t)$ at time t and distance r from the centre of the symmetry, c being the speed of sound. There are two basic types of solutions: expanding or “retarded” waves

$$\phi = \frac{1}{r} f(r - ct) \quad (25)$$

for a suitable function $f(x)$, and contracting or “advanced” waves

$$\phi = \frac{1}{r} f(r + ct). \quad (26)$$

A solution of the retarded type might describe sound waves emanating from a whistle at the centre of symmetry. On the other hand, solutions of the advanced type are generally an embarrassment to the theory, as they describe spherical waves converging coherently on the centre of symmetry. Note how time reversal interchanges the retarded and advanced waves. Invocation of the Second Law of Thermodynamics as a selection principle excludes advanced waves as being improbable. However, in the presence of sufficiently sophisticated apparatus, advanced waves as in (26) may actually occur. One example is given by the firing of the detonation lens of an atomic bomb, clearly a very special arrangement. Another example occurs at one focus of an ellipsoidal (or semi-ellipsoidal) “whispering gallery,” if sound waves are emanating from a “whisperer” located at the other focus. For this situation to arise, the ceiling of the gallery has to be specially designed and carefully constructed.

The so-called “Gibbs Paradox” may be resolved by similar considerations of the power of the apparatus involved. Consider a chamber containing two different gases. Configurations with one gas entirely on one side of the chamber and the other gas on the other side are clearly improbable. If the available apparatus is able to detect the difference between the two gases, then it could extract work along with the thermodynamic entropy gain that would ensue as the gases mixed together. On the other hand, if the apparatus was not sophisticated enough to detect the difference between the two gases, it could not distinguish separated configurations from those in which the two gases were mixed indiscriminately. Under these conditions, the thermodynamic entropy of a separated configuration would be the same as the thermodynamic entropy of a mixed state.

5 Continuous distributions

Equation (1) represents the information-theoretic entropy of the discrete probability distribution $\xi = (p_1, \dots, p_r)$. In the literature, one often encounters an extension of (1) to the case of a one-dimensional random variable X with density function $p(x)$ in the form

$$H(X) = - \int_{-\infty}^{\infty} p(x) \log p(x) dx \quad (27)$$

[e.g §8.3 of Ash (1965), p.541 of Rumer and Ryvkin (1977), §20 of Shannon (1948)]. There is nothing wrong with (27) as a pure-mathematical formula (assuming convergence of the integral and absolute continuity of the density p with respect to Lebesgue measure). However, in physical applications, the coordinate x in (27) represents an abscissa, a distance from a fixed reference point. This distance x has the dimensions of length. The density function $p(x)$ is specified so that

$$P(a \leq X < b) = \int_a^b p(x) dx \quad (28)$$

is the probability that the random variable X falls within the interval $[a, b)$. As a probability, the left hand side of (28) is dimensionless. Since the infinitesimal dx on the right hand side has the dimensions of length, it follows that the density $p(x)$ has the dimensions of $(\text{length})^{-1}$. Now for $0 \leq z < 1$, one has the series expansion

$$-\log(1 - z) = z + \frac{1}{2}z^2 + \frac{1}{3}z^3 + \dots \quad (29)$$

For consistency in (29), it is necessary that the argument of a logarithm be dimensionless. The formula (27) is then seen to be dimensionally incorrect, since the argument of the logarithm on its right hand side has the dimensions of a probability density. [Interestingly enough, although Shannon (1948) uses the formula (27), he does note its lack of invariance with respect to changes in the coordinate system, and shows how to transform the formula (27). The dimensional incorrectness of Shannon’s formula was also noted by O’Neill (1963), using a somewhat different argument to the one given above.]

In order to make the continuous entropy definition (27) consistent, one has to normalize the argument of the logarithm by dividing $p(x)$ by another density. Let this density be the *reference density* $q(x)$. Formula (27) then becomes

$$-\int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx. \quad (30)$$

Now if η is a probability distribution (q_1, \dots, q_r) , one defines the “cross-entropy” or “Kullback entropy” or *entropy of ξ relative to η* as

$$H(\xi|\eta) = -\sum_{i=1}^r p_i \log \frac{p_i}{q_i} = \sum_{i=1}^r p_i (\log q_i - \log p_i) \quad (31)$$

in the finite case (Kullback, 1959). Note that if η is the entropy-maximizing uniform distribution or “microcanonical” distribution μ with $q_i = 1/r$ for $1 \leq i \leq r$, then (31) reduces to

$$H(\xi|\mu) = -\log r + H(\xi), \quad (32)$$

a non-positive quantity by (10). Maximization of $H(\xi)$ is equivalent to maximization of (32).

Returning to the continuous case, suppose that the reference density determines the distribution of a random variable Y according to the analogue

$$P(a \leq Y < b) = \int_a^b q(x) dx \quad (33)$$

of (28). One may then interpret the dimensionally correct quantity (30) as the entropy

$$H(X|Y) = -\int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \quad (34)$$

of X relative to the *reference random variable* Y . There remains the problem of choosing an appropriate reference random variable. From a Bayesian standpoint, Jaynes (1963) and Garrett (1991) discuss the use of an “improper prior”, taking $q(x)$ to be a Lebesgue measure justified by symmetry considerations. This prior is “improper” to the extent that Lebesgue measure is not normalizable to a probability density, so it does not correspond to a reference random variable Y . A more satisfactory choice of reference density follows from physical considerations. Since one only has a finite time t in which to carry out an experiment, one can only scan an interval of length $2ct$, where c is the speed of light. (In higher dimensions, this length should be replaced by an appropriate volume, such as $\frac{4}{3}\pi c^3 t^3$ in three dimensions.) One knows that no information can be extracted during the experiment from events outside this window. Thus the most random density subject to the given constraint is the density that is uniform inside the window, and zero outside. This is a true (i.e. normalizable) density, corresponding to a proper reference random variable M . With this choice, (34) reduces to

$$H(X|M) = - \int_{x_0-ct}^{x_0+ct} p(x) \log 2ctp(x) dx, \quad (35)$$

where x_0 is the location of the scanning device.

6 Conclusion

The entropy concept, be it information-theoretic or thermodynamic, depends critically on the nature of the apparatus used. In apparently pure mathematical formulae such as (1), this dependence is built in to the definition of the underlying probabilities. Awareness of the dependence on the experimental setup allows one to avoid many of the paradoxes associated with the entropy concept. It also yields a well-defined and dimensionally correct entropy for continuous distributions, namely the entropy (35) relative to a reference distribution representing the maximum information obtainable within the duration of an experiment.

References

- [1] Ash, R.B. (1965). *Information Theory*, Interscience, New York, NY.
- [2] Feynman, R.P., R.B. Leighton, and M. Sands (1963). *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley, Reading, MA.
- [3] Garrett, A.J.M. (1991). Macroirreversibility and microreversibility reconciled, in *Maximum Entropy in Action* (eds. Buck, B. and V.A. Macaulay), 139, Clarendon Press, Oxford.
- [4] Jaynes, E.T. (1963). Information theory and statistical mechanics, in *Statistical Physics*, 1962 Brandeis Lectures (ed. Ford, K.W.), 181, Benjamin, New York, NY.
- [5] Kullback, S. (1959) *Information Theory and Statistics*, Wiley, New York, NY.
- [6] Li, M. and P. Vitanyi (1997) *An Introduction to Kolmogorov Complexity and its Applications*, Springer, New York, NY.
- [7] Martin-Löf, P. (1966). The definition of random sequences. *Information and Control* 9: 602.
- [8] O'Neill, E.L. (1963). *Introduction to Statistical Optics*, Addison-Wesley, Reading, MA.
- [9] Pauli, W. (1933). *Handbuch der Physik* xxiv/1: 151.
- [10] Prigogine, I. and C. George (1983). The Second Law as a selection principle: the microscopic theory of dissipative processes in quantum systems. *Proc. Nat. Acad. Sci.* 80: 4590.
- [11] Rumer, Yu.B. and M.Sh. Ryvkin (1977). *Termodinamika, Statisticheskaya Fizika i Kinetika*. English Translation (1980): *Thermodynamics, Statistical Physics, and Kinetics*, Mir, Moscow.
- [12] Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27: 379. Reprinted in Shannon, C.E. and W. Weaver (1949): *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.
- [13] Tolman, R.C. (1938). *The Principles of Statistical Mechanics*, Oxford University Press, Oxford.
- [14] Uspenskiĭ, V. A., A.L. Semenov, and A.Kh. Shen' (1990). Can an (individual) sequence of zeros and ones be random? (Russian) *Uspekhi Mat. Nauk* 45: 105. English Translation: *Russian Mathematical Surveys* 45 : 121.