NC107 April 16, 1962
Physics

IBM Research Report

A Critical Note on Information Theory and Statistical Mechanics

Benoit Mandelbrot
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Thomas J. Watson Research Center, Yorktown Heights

A CRITICAL NOTE ON INFORMATION THEORY AND STATISTICAL MECHANICS*

by

Benoit Mandelbrot
4/16/62

ABSTRACT: This is an examination of the central arguments of a number of papers, which attempt to derive the canonical distribution of energy (due to Gibbs) from a principle of maximization of Shannon's information. We stress that there are many other concepts of information, so that the idea of information maximization is as ambiguous as the other parts of the theory of inductive behavior. Besides, we stress that the "information" approach does not allow one to dispense with a postulate equivalent to the second principle of phenomenological thermodynamics.

---

*This work was partly supported by the Office of Naval Research, under contract number Nonr-3775(00), NR-047040.

IBM

LIMITED DISTRIBUTION NOTICE

This report has been submitted for publication elsewhere and has been issued as a Research Note for early dissemination of its contents. As a courtesy to the intended publisher, it should not be widely distributed until after the date of outside publication.

Printed in U.S.A.

1. Introduction

Recent discussions of the relations between the concepts of entropy, "negentropy" and information, have revived the controversy and the irritation which seem always to accompany the discussion of the role of the observer in the structure of statistical thermodynamics. Among the authors who believe that information should play a role in thermodynamics, two main viewpoints can be distinguished:

Some authors consider that the structure of the "special" thermodynamical theory of equilibrium has been well established without any reference to the concept of information. They use the latter only to extend the scope of thermodynamical arguments, not denying the arbitrariness which is introduced with the observer. This was the case of Szilard (11b and 11c), who had earlier given his own set of thermodynamical foundations (11a). Similar attitudes can be found in the work of Brillouin (1) and in our own (9a to 9d).

Other authors believe on the contrary that the concept of information should be used in the very foundation of even the most classical parts of thermodynamics. A particularly forceful and well-written exposition of this approach has been given by E. T. Jaynes (6); it was adopted to engineering instruction by M. Tribus (12) and was extended by R. Ingarden and K. Urbanik (5). In P. T. Landsberg's recent treatise (8), a similar set of calculation is reinterpreted in an "objective" way, but it does not play any central role in the development.

The primary purpose of the present Note is to examine what has come to be called "Jaynes' method" and to show that its formal simplicity and its seeming "obviousness" are both misleading.

First of all, Jaynes' approach seems so simple because it neglects to stress one of its fundamental axioms; although the concept of entropy is basic to his approach and serves to derive the canonical law, Jaynes' axioms are not sufficient to show that the temperature which occurs in the canonical law is the same as the integrating divisor of "heat". His method for proving this essential point is traditional in statistical mechanics and —as usual—he does not stress the assumption that the discrete "states" of a physical system are neither created nor destroyed when the external parameters are varied. (This is the simplest case of what is called generally the adiabatic invariance of the weight of degeneracy). Of course, this conservation of "states" seems to be a most natural assumption; but it only means that a well-chosen term can go a long way towards making a postulate look obvious while it is not. In statistical mechanics, one must prove that "state" is more than a word; in thermodynamics, one can prove that "adiabatic invariance" is equivalent to the second principle of thermodynamics in Clausius' or Carathéodory's forms. If this additional principle is stressed, Jaynes' approach becomes much less simple.

Second, Jaynes' approach seems obvious because it relies upon Shannon's proof of the unicity of his concept of "information". However, the circumstances encountered in physics destroy this unicity and special (less intuitive) arguments are necessary in order to show, for example, that "information" is not the concept bearing the same name which was introduced in 1925 by R. A. Fisher, or the concept of information of Wald-Kullback-Leibler, or perhaps some other version of the general concept of information introduced by M. P. Schutzenberger (10a and 10b). Hence, the method based upon a maximization of Shannon's information can carry no more conviction than a host of other more or less arbitrary inductive procedures.

Note also that Jaynes' approach is too successful in explaining what is "the" entropy. That concept has "so many faces" to use a phrase of H. Grad (4), that it cannot be represented — even in the case of equilibrium — by any single mathematical formula.

We shall say no more about the questions related to the second principle, and shall rather give more details about information-maximization, which turns out to be related to the zeroeth principle. We shall make no pretense of being a disinterested party and shall frequently refer to (11a and 11b) and to (9a to 9d).

It should be stressed that we have no quarrel with either of the two well-known interpretations of thermodynamics; the subjectivistic one — of which Jaynes gives a good if somewhat extreme exposition — and the objectivistic one — which is used by Landsberg. We think that it is most desirable that a foundation for thermodynamics be interpretable in both ways.

2. The Canonical Distribution and Baltzmann's Derivation

A central point of statistical thermodynamics is the canonical distribution of Gibbs, which we shall write as follows:

$$\Pr(\text{any state of energy } u) = \exp(-\beta u)/Z(\beta),$$
$$\Pr(u \leq \text{energy} \leq u + du) = dG(u)\exp(-\beta u)/Z(\beta),$$

where G is the number of states of energy not greater than u.

One of the basic derivations of this distribution is a generalization by Einstein of an argument due to Boltzmann.

After many approximations, one shows the following: $p(r)$ being the probability of encountering a partial system S_m in the state Q_r, one should expect an isolated system $S = \Sigma S_m$, of energy u, to be in the state in which one attains the maximum of

$$-\Sigma p(r) \log p(r), \quad \text{given} \quad \Sigma p(r) = 1 \quad \text{and} \quad \Sigma p(r) u_r = u.$$

(This property is derived by Gibbs as a theorem concerning the canonical distribution).

Unfortunately, this characterization involves a succession of approximations, the scope of which is difficult to evaluate. In particular, the expression $-\Sigma p(r) \log p(r)$ plays such a central role in the theory, that one would like to derive it more directly. Moreover, in carrying out the maximization of $-\Sigma p(r) \log p(r)$, one must replace u by $E(U)$; this operation is usually performed too casually, as we have stressed in (9), and it may be preferable to motivate directly both the idea of maximization and the choice of quantity to maximize.

3. Use of the Axiomatics of Shannon's Information

The striking formal analogy between the entropy and Shannon's information has suggested two kinds of reinterpretations of the operation

$$\text{maximize} \quad -\Sigma p(r) \log p(r), \quad \text{given} \quad \Sigma p(r) = 1 \quad \text{and} \quad \Sigma p(r) u(r) = E(U).$$

Objective Approach: Step One

The concept of entropy is borrowed from non-statistical thermodynamics and it is assumed that, at equilibrium, it attains the maximum compatible with all the other constraints.

Step Two

One assumes that entropy must a priori satisfy certain axioms, which turn out to be also true of Shannon's information. Then, according to a theorem of Shannon, entropy must be represented by the expression

"$-\Sigma p(r) \log p(r)$" which must be maximum in the state of equilibrium.

Subjective Approach: Step One

The concept of the "least-prejudiced" inductive procedure is introduced as being the best way of behaving in the presence of uncertainty.

Step Two

One assumes that "prejudice" must a priori satisfy certain axioms, which turn out to be also true of Shannon's information. Then, according to a theorem of Shannon, prejudice must be measured by information, and "$-\Sigma p(r) \log p(r)$" must again be maximized.

4. A Critique of Step One

We have, of course, nothing against the process of maximization, which—once the foundations of the theory have been laid—frequently provides the most direct way of going further.[1] But, in our opinion, maximization is inacceptable in the foundations, for the following reasons.

In the objectivistic interpretation, it postulates that non-statistical thermodynamics has already been established, and that one wants to generalize its principles to obtain a fuller, statistical, description. As a matter of fact, in order to base the statistical thermodynamics of equilibrium, this procedure requires a definition of entropy so general, that it applies to non-equilibrium situations as well. Under these conditions, one cannot fulfill one of the aims of the statistical theory: to derive the non-statistical approach as a theory of expected behaviors. This unsatisfactory

[1] Note that our comments concerning maximization of information do not apply to our work on the statistical properties of wordcounts. In that case, information is not maximized because it has certain axiomatic properties, but because it is a central concept in the theory of coding of messages. See the end of Section 5.

situation becomes quite untolerable if one's ambitions are increased, and if—following (11) or (9)—one wishes to have a foundation of thermodynamics that is immediately both phenomenological and statistical.

In its subjectivistic interpretation, the approach based upon maximization stands or falls with a certain method of inductive behavior, which happens not to have benefited so far from the energetic discussion to which statisticians usually subject proposals of this kind. [The only discussion that we can mention here is due to the physicist Cox (2)]. Of course, the statisticians' fashions are not to be followed blindly. But, in view of their well-known inadequacies, even the "best" methods of induction cannot be welcomed in the foundations of thermodynamics. The situation becomes—again—untolerable when one notes that the canonical distribution of Gibbs also happens to be "optimal" from the viewpoints of a host of other inductive procedures. As a matter of fact, we saw in (9a) that "Gibbs' law" is a synonym for the statisticians' "distribution of the exponential type" and that it has the property of "statistical sufficiency" that makes it optimal from the viewpoint of almost every inductive procedure. Since sufficiency can be interpreted subjectively, the stress upon any special inductive procedure is conceptually misleading. Under the circumstances, one cannot put any trust in an argument that singles out the method based upon information maximization.

[However, it is clear that the arguments using "statistical sufficiency" are far less obvious than those based upon entropy maximization. Hence, from the viewpoint of pedagogy, there may be advantage in the following roundabout procedure: A) Argue that—in order for statistical sufficiency to hold—it is necessary that any specific procedure—such as maximum likelihood estimation of maximum information behavior—depend only upon the total energy of an isolated system. B) Derive the canonical law in this way. C) Verify a posteriori that the canonical law is logically sufficient as well as necessary for statistical sufficiency to hold.]

Actually, the information-theoretical approach again yields more than the canonical distribution, because it defines entropy for all non-canonical distributions. This procedure may be necessary in non-equilibrium thermodynamics; but this is still a most questionable field and it seems to be poor pedagogy, to require it in order to derive the wholly unquestioned equilibrium theory.

5. A Critique of Step Two

Many professional statisticians expressed surprise at Shannon's axiomatic of information, since they had long before used another concept, due to R. A. Fisher, which was also based upon an (informal) axiomatic of the common idea of "information".[2] Wald [see (7)] uses still

---

[2] On p. 47 of essay number 26 of (3), Fisher writes: "... The amount of information is calculable... In introducing the concept of quantity of information we do not want merely to be giving an arbitrary name to a calculable quantity, but must be prepared to justify the term employed, in relation to what common sense requires, if the term is to be appropriate, and serviceable as a tool for thinking. The mathematical consequences of identifying, as I propose, the intrinsic accuracy of the error curve, with the amount of information extracted, may therefore be summarized specifically in order that we may judge by our pre-mathematical common sense whether they are the properties it ought to have.

First, then, when the probabilities of the different kinds of observation which can be made are all independent of a particular parameter, the observations will supply no information about the parameter. Once we have fixed zero we can in the second place fix totality. In certain cases estimates are shown to exist such that, when they are given, the distributions of all other estimates are independent of the parameter required. Such estimates, which are called sufficient, contain, even from finite samples, the whole of the information supplied by the data. Thirdly, the information extracted by an estimate can never exceed the total quantity present in the data. And, fourthly, statistically independent observations supply amounts of information which are additive. One could, therefore, develop a mathematical procedure. It is, perhaps, only a personal preference that I am more inclined to examine the quantity as it emerges from mathematical investigations, and to judge of its utility by the free use of common sense, rather than to impose it by a formal definition".

another concept of information, which has also come to play a big role in the work of the Russian school of probability. The confusion was increased by Wiener's casual remark that the Shannon-Wiener information can be substituted to that of Fisher. Finally, various students of human organizations and of inductive behavior have used still other concepts of the same name, semantic or otherwise.

There is, therefore, surely no concensus of opinion about the identification of physical entropy with Shannon's information. As a matter offact, Fisher found a relation between entropy and *his* information.[3]

This forces a re-examination of the axiomatics of Shannon's information, and one sees indeed that these axioms are not necessarily relevant to physics. (For example, in order to give a physical meaning to the axiom referring to equiprobable events, it is necessary to consider systems of fixed energy, having $dG(u)$ equi-probably "states". But, if two such systems are put together into one, $G(u) = \int G'(u')\, dG''(u - u')$ so that the information $\log (dG)$ of the whole is *not* the sum of the informations $\log(dG')$ and $\log(dG'')$ of the two parts; hence, the axiom of additivity refers to a physically non-realizable situation).

A more fundamental question is the following: Shannon's derivation assumes that $p(r)$ does not depend upon any extraneous parameter. But—after the expression for information has been derived—it is immediately

---

[3] The quotation of footnote (2) continues as follows: ..."As a mathematical quantity information is strikingly similar to entropy in the mathematical theory of thermodynamics. You will notice especially that reversible processes, changes of notation, amthematical transformations if single-valued, translation of the data into foreign languages, or rewriting them in code, cannot be accompanied by loss of information; but that the irreversible processes involved in statistical estimation, where we cannot reconstruct the original data from the estimate we calculate from it, may be accompanied by a loss, but never by a gain".

applied to the derivation of a distribution parametrized by temperature. Actually, if one puts the parameter in at the beginning, and if one uses a somewhat more general set of axioms, due to M. P. Schutzenberger (10), one may obtain any one of the expressions

$$-\Sigma p(r|\theta)\, L\, \{\log [p(r|\theta)]\},$$

where L is some linear operator. For example, Wald's information corresponds to the shift operator relative to a parameter, that is to $L[G(\theta)] = G(\theta) - G(\theta^{*})$; Fisher's information corresponds to $L[G(\theta)] = (\partial^2/\partial\theta^2)\, G(\theta)$; and Shannon's information corresponds to $L[G(\theta)] = \text{constant}.\ G(\theta)$.

As a result, Shannon's information becomes again unique if one adds the condition that there are no outside parameters. Alternatively, one may add as an axiom an innocuous-looking condition of normalization. In other words, in order to make Shannon's information unique, one must: a) <u>either</u> make the axiomatic more stringent than that of Schutzenberger (some of the axioms will then be of misleading inocuity); b) <u>or</u> take account of the fact that, in the "special" case of canonical systems, the concept of entropy can be introduced on the basis of phenomenological statistical principles that do not suffer from undeterminate inference; therefore, an information-theoretical generalization of entropy must coincide with the ordinary concept for canonical systems, and this determines information as being Shannon's.

Very similar observations have long ago been made in the original context of Shannon's information, namely the theory of the transmission of digital data. The main fact there is the existence of inequalities in which one side is the something interpreted as "information". The axiomatic approach is only a dressing that makes certain words palatable; it has turned out to be a mistake to over-emphasize it.

6. Conclusion

Even if a foolproof axiomatic of "entropy-information" were available, the argument using this concept has none of the obviousness that it seems to enjoy. Of course, the ultimate criterion in this whole question is one of pedagogy; but—if one insists upon using information—the best is to introduce it heuristically and not to skim it too much.

REFERENCES

1. Brillouin, Leon: Science and Information Theory, New York, Academic Press (1956), (Second Edition, 1961).

2. Cox, Richard T.: The Algebra of Probable Inference, Baltimore: The Johns Hopkins Press (1961).

3. Fisher, R. A.: Contributions to Mathematical Statistics, New York, Wiley (1950).

4. Grad, Harold: "The Many Faces of Entropy", Comm. Pure and Appl. Math. 14 (1961), 323-354.

5. Ingarden, Roman S. and Urbanik, K.: "Quantum Informational Thermodynamics", Acta Physica Polonica, Vol. 21 (1962) 281-304.

6. Jaynes, E. T.: "Information Theory and Statistical Mechanics", The Physical Review, Vol. 106 (1957), 620-230 and Vol. 108 (1957), 171-190.

7. Kullback, Solomon: Information Theory and Statistics, New York, Wiley (1959).

8. Landsberg, T. T.: Thermodynamics, New York, Intersicence (1961).

9. Mandelbrot, Benoit: "The Role of Sufficiency and of Estimation in Thermodynamics", Annals of Mathematical Statistics, 33 (1962); b) "Derivation of Statistical Thermodynamics from Purely Phenomenological Principles", IBM Research Report NC-106; c) "Exhaustivite de l'Energie Totale d'un Systeme, Pour l'Estimation de sa Temperature", Comptes Rendus de l'Academie des Sciences de Paris, Vol. 243 (1956), 1835-1838; d) "An Outline of a Purely Phenomenological Theory of Statistical Thermodynamics",

Institute of Radio Engineers (IRE) Transactions on Information Theory, Vol. IT-2 (Sept. 1956), 190-203.

10. Schutzenberger, M. P.: a) "Sur les Rapports Entre la Quantite d'Information au Sens de Fisher et au Sens de Wiener", Comptes Rendus de l'Academie des Sciences de Paris, Vol. 232 (1951), 925-927; b) "Some Measures of Information Used in Statistics" Information Theory-the Third London Symposium (Edited by Colin Cherry) New York, Academic Press (1956), 18-25.

11. Szilard, Leo: "Uber die Ausdehnung der Phänomenologischen Thermodynamik auf die Schwankungserscheinungen", Z. Phys. Vol. 32 (1925), 753-788; b) "Uber die Entropieverminderung in Einem Thermodynamischen System bei Eingriff Intelligenter Wesen", Z. Phys. Vol. 53 (1929), 840-856.

12. Tribus, Myron: Thermostatistics and Thermodynamics, Princeton, van Nostrand (1962).