Journal of Statistical Physics, Vol. 1, No. 3, 1969

A New Theorem of Information Theory

Arthur Hobson[1]

Received September 9, 1969

Consider a random experiment whose possible outcomes are z_1, z_2,..., z_n. Let the prior probabilities be p_1^0,..., p_n^0, and let the posterior probabilities be p_1,..., p_n. It is shown that, subject to certain prescribed and intuitively reasonable conditions, the expression $I = k \sum p_i \ln(p_i/p_i^0)$, where k is a positive constant, is the unique expression for the information contained in a message which alters the probabilities from the p_i^0 to the p_i.

KEY WORDS: information theory; mathematics; probability theory; quantitative measure of uncertainty; communication theory; entropy; foundations of statistical mechanics.

1. INTRODUCTION

Information theory was founded in 1948 when Shannon[1] introduced the useful quantitative measure

$$U(p_1, p_2, ..., p_n) = -k \sum p_i \ln p_i \tag{1}$$

for the "uncertainty" or "missing information" in the discrete probability distribution $(p_1, ..., p_n)$. One reason for the wide use of (1) is that it is the *only* expression possessing certain intuitively reasonable properties which are desirable in an uncertainty measure. That is, Shannon proved[1] a uniqueness theorem to the effect that, *if* the function $U(p_1, ..., p_n)$ possesses certain intuitively reasonable properties, *then* this function is necessarily given by (1). Shannon's uniqueness theorem has been subsequently reproven in various ways by several authors (see, for instance, the elegant proofs of Feinstein[2] and Khinchin[3]).

It has been found, however, that it is hard to generalize (1) to continuous probability distributions $f(x)$.[4,5] The problem is that any reasonable generalization yields $U = \infty$, which is not very useful. Furthermore, if the generalization is carried out

[1] Department of Physics, University of Arkansas, Fayetteville, Arkansas.

in the most straightforward way and the infinity is then "subtracted" (such a subtraction is, of course, nonrigorous), the resulting expression

$$U[f(x)] = -k \int f(x) \ln f(x)\, dx \tag{2}$$

is not invariant under a change of variables $x \to y = y(x)$. The difficulty with (1) may be stated more precisely as follows: It is generally possible to treat probability theory within a single mathematical formalism, valid for both discrete and continuous distributions, by the use of Lebesgue–Stieltjes integrals. However, there is no single Lebesgue–Stieltjes integral which reduces to (1) in the discrete case and to (2) in the continuous case. Thus, so long as (1) and (2) are used as the basic information measures, information theory must always treat the discrete and continuous cases by means of two different mathematical formalisms.

In order to escape the difficulties associated with (1), several authors[6–9] have proposed that the basic information measure should instead be

$$I(p_1,\ldots,p_n\,;p_1^0,\ldots,p_n^0) = k \sum p_i \ln(p_i/p_i^0) \tag{3}$$

where $(p_1,\ldots,p_n)$ and $(p_1^0,\ldots,p_n^0)$ are two discrete probability distributions. The interpretation of (3) will be given in the next section. This expression seems to have all the properties desired in an information measure; in fact, in a certain sense, (3) provides a generalization of Shannon's expression (1).[6,8,10] Note, however, that (3) is *not* simply the difference between $U(p_1^0,\ldots,p_n^0)$ and $U(p_1,\ldots,p_n)$.

Expression (3) does not suffer from the difficulties associated with (1). It is easy to generalize (3) to continuous distributions $f(x)$ and $f^0(x)$, and the resulting expression

$$I[f(x);f^0(x)] = k \int f(x) \ln[f(x)/f^0(x)]\, dx \tag{4}$$

does not exhibit divergence or invariance difficulties.[10] Both (3) and (4) may be encompassed within a single Lebesgue–Stieltjes integral expression.[6,8]

As has been shown by Schlögl,[9] and as is further demonstrated elsewhere,[10] expression (3) is of fundamental importance in statistical mechanics.

Despite the appealing properties of (3), it has never been clear that (3) is the *only* expression having these properties. That is, a uniqueness theorem for (3), similar to Shannon's uniqueness theorem for (1), has not been given. The purpose of this paper is to state and prove such a theorem.

2. STATEMENT OF THE THEOREM

Consider an observer who is interested in some experiment having possible outcomes $z_1,\ldots,z_n$, but who does not know the precise outcome z_i. Let his initial "state of knowledge" about the outcome be described by the prior probabilities $p_1^0,\ldots,p_n^0$. Now, suppose that our observer is handed a message containing new data about the outcome, and that the observer's state of knowledge after receiving the message is described by the posterior probabilities $p_1,\ldots,p_n$. We seek an expression for the amount of information I in the message.

The information should be a function $I(p_1,..., p_n ; p_1^0,..., p_n^0)$ of the prior and posterior probabilities. It seems reasonable to postulate that

$$I \quad \text{is a continuous function of its} \quad 2n \quad \text{variables,} \tag{5}$$

and

$$I(p_1,..., p_j,..., p_k,..., p_n ; p_1^0,..., p_j^0,..., p_k^0,..., p_n^0)$$
$$= I(p_1,..., p_k,..., p_j,..., p_n ; p_1^0,..., p_k^0,..., p_j^0,..., p_n^0) \tag{6}$$

and

$$I = 0 \qquad \text{if} \quad p_i = p_i^0 \quad \text{for all} \quad i \tag{7}$$

Postulate (5) says that the information is changed only slightly when the probabilities are changed only slightly; (6) says that the information does not depend on the order in which the possibile outcomes z_i are labeled; (7) says that no information is obtained if the message results in no change in the probabilities.

For any pair of integers n, n_0 such that $n_0 \geqslant n$, the expression $I(1/n,..., 1/n, 0,..., 0; 1/n_0,..., 1/n_0)$ represents the information obtained when the number of equally likely possibilities is reduced from n_0 to n. It is reasonable to postulate that

$$I(1/n,..., 1/n, 0,..., 0; 1/n_0,..., 1/n_0) \quad \text{is an increasing function of} \quad n_0 \quad \text{and a decreasing function of} \quad n, \text{ for any integers} \quad n, n_0 \quad \text{such that} \quad n_0 \geqslant n \tag{8}$$

This says that the information is greater when the number n_0 of prior possibilities is greater and/or the number n of posterior possibilities is smaller.

Finally, a condition known as the "composition rule" will be needed. In probability theory, the set $S \equiv \{z_1, z_2,..., z_n\}$ of possible outcomes is called a "probability space," and any subset $E \subset S$ is called an "event." Suppose that S is divided into two events $E_1 = \{z_1,..., z_r\}$ and $E_2 = \{z_{r+1},..., z_n\}$. The posterior probabilities of these events are

$$P(E_1) \equiv q_1 = p_1 + \cdots + p_r, \qquad P(E_2) \equiv q_2 = p_{r+1} + \cdots + p_n \tag{9}$$

while the prior probabilities of E_1 and E_2 are

$$P^0(E_1) \equiv q_1^0 = p_1^0 + \cdots + p_r^0, \qquad P^0(E_2) \equiv q_2^0 = p_{r+1}^0 + \cdots + p_n^0 \tag{10}$$

The posterior probabilities of the z_i, given that E_1 occurred, are

$$P(z_i \mid E_1) = p_i/q_1 \qquad (i = 1,..., r) \tag{11}$$

while the prior probabilities of the z_i, given E_1, are

$$P^0(z_i \mid E_1) = p_i^0/q_1^0 \qquad (i = 1,..., r) \tag{12}$$

Similarly,

$$P(z_i \mid E_2) = p_i/q_2 \qquad (i = r + 1,..., n) \tag{13}$$

$$P^0(z_i \mid E_2) = p_i^0/q_2^0 \qquad (i = r + 1,..., n) \tag{14}$$

Now, information about the outcome may be given *either* by specifying the probabilities $p_1,..., p_n$ directly, *or* by specifying the probabilities (9) of the two events

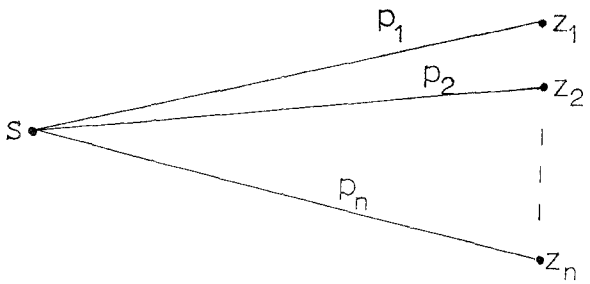

Fig. 1. A representation of the probability distribution $p_1, \ldots, p_n$ on the space $S = \{z_1, \ldots, z_n\}$.

E_1 and E_2 and then giving the conditional probabilities (11) and (13). These two alternatives are shown diagrammatically in Figs. 1 and 2. We now postulate

$$I(p_1, \ldots, p_r, p_{r+1}, \ldots, p_n ; p_1^0, \ldots, p_r^0, p_{r+1}^0, \ldots, p_n^0)$$
$$= I(q_1, q_2 ; q_1^0, q_2^0) + q_1 I(p_1/q_1, \ldots, p_r/q_1 ; p_1^0/q_1^0, \ldots, p_r^0/q_1^0)$$
$$+ q_2 I(p_{r+1}/q_2, \ldots, p_n/q_2 ; p_{r+1}^0/q_2^0, \ldots, p_n^0/q_2^0) \tag{15}$$

Postulate (15) says that the amount of information in the scheme represented by Figure 1 equals the amount of information in the equivalent scheme of Figure 2, and, furthermore, that the right member (15) is a reasonable expression for the

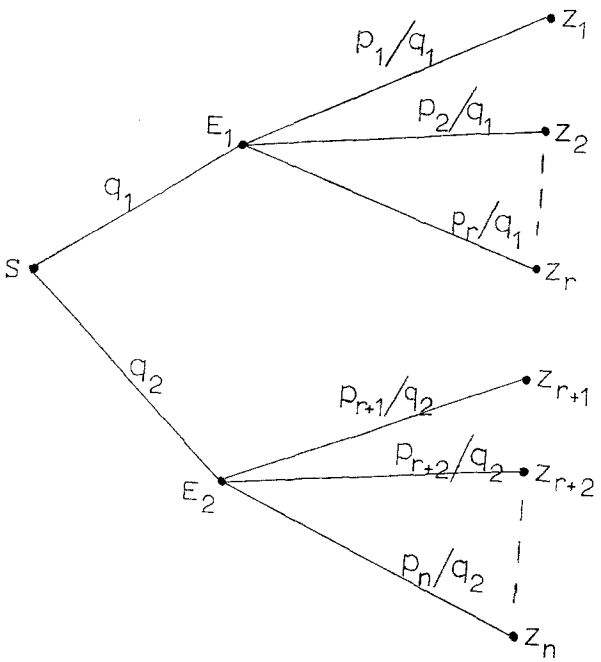

Fig. 2. An alternative way of representing the probability distribution $p_1, \ldots, p_n$. The symbols q_1 and q_2 are defined by (9).

information in the scheme of Figure 2. Thus, the basic idea of the composition rule is the assumption that the information in the scheme of Figure 2 is the *sum* of the information $I(q_1, q_2; q_1^0, q_2^0)$ concerning the first step (i.e., concerning which of the events E_1 or E_2 occurred) plus the weighted [with weighting factors $P(E_1)$ and $P(E_2)$] information concerning the second step. This seems reasonable if one thinks about it awhile; at least, the right member of (15) seems more reasonable than any *other* expression for the information in the scheme of Figure 2.

Uniqueness Theorem. Let $I(p_1, \ldots, p_n; p_1^0, \ldots, p_n^0)$ be defined for any pair of probability distributions $(p_1, \ldots, p_n)$ and $(p_1^0, \ldots, p_n^0)$ on a finite probability space $S = \{z_1, \ldots, z_n\}$, and let I satisfy the postulates (5)–(8) and (15). Then I is necessarily given by (3), where k is a positive constant.

Thus, (3) is the *only* expression having the intuitively reasonable properties (5)–(8) and (15).

Shannon's uniqueness theorem for the expression (1) is similar to the above uniqueness theorem for (3). However, Shannon's theorem (and its proof) is somewhat less complicated than the present theorem, since (1) involves only a single probability distribution, whereas (3) involves two distributions. Portions of the present proof will follow the proofs of Shannon's theorem given by Feinstein[2] and Khinchin.[3]

3. PROOF OF THE THEOREM

We first show that[2]

$$I(q_1, \ldots, q_{n-1}, q_n; ---) = I(q_1, \ldots, q_{n-2}, q_{n-1} + q_n; ---) + (q_{n-1} + q_n) I\left(\frac{q_{n-1}}{q_{n-1} + q_n}, \frac{q_n}{q_{n-1} + q_n}; ---\right) \tag{16}$$

Proof of (16). By (15), with $Q \equiv q_1 + \cdots + q_{n-2}$,

$$I(q_1, \ldots, q_n; ---) = I(Q, q_{n-1} + q_n; ---) + QI\left(\frac{q_1}{Q}, \ldots, \frac{q_{n-2}}{Q}; ---\right) + (q_{n-1} + q_n) I\left(\frac{q_{n-1}}{q_{n-1} + q_n}, \frac{q_n}{q_{n-1} + q_n}; ---\right) \tag{17}$$

Again using (15),

$$I(q_1, \ldots, q_{n-2}, q_{n-1} + q_n; ---) = I(Q, q_{n-1} + q_n; ---) + QI(q_1/Q, \ldots, q_{n-2}/Q; ---) + (q_{n-1} + q_n) I(1; 1) \tag{18}$$

By (7), $I(1; 1) = 0$. Combining (17) and (18), we obtain (16).

[2] Concerning the notation: Dashes appearing to the right of the semicolon will always mean that the variables on the right are identical with those on the left, only with the superscript "0" attached. For instance, $I(p_1, p_2; ---)$ means $I(p_1, p_2; p_1^0, p_2^0)$.

Next, we show that

$$I(p_{11},\ldots,p_{1m},\text{---},p_{n1},\ldots,p_{nm};\text{---}) = I(q_1,\ldots,q_n;\text{---}) + \sum_{i=1}^{n} q_i I(p_{i1}/q_i,\ldots,p_{im}/q_i;\text{---}) \tag{19}$$

where $q_i \equiv p_{i1} + \cdots + p_{im}$.

Proof of (19). By (15), Eq. (19) holds for $n = 2$. We proceed by induction. Assuming (19) holds for n,

$$\begin{aligned} I(p_{11},\ldots,p_{1m},\text{---},p_{n+1,1},\ldots,p_{n+1,m};\text{---}) &= I(q_1,\ldots,q_{n-1},q_n+q_{n+1};\text{---}) \\ &+ \sum_{i=1}^{n-1} q_i I\left(\frac{p_{i1}}{q_i},\ldots,\frac{p_{im}}{q_i};\text{---}\right) \\ &+ (q_n+q_{n+1})\, I\left(\frac{p_{n1}}{q_n+q_{n+1}},\ldots,\frac{p_{nm}}{q_n+q_{n+1}},\frac{p_{n+1,1}}{q_n+q_{n+1}},\ldots,\frac{p_{n+1,m}}{q_n+q_{n+1}};\text{---}\right) \end{aligned} \tag{20}$$

By (16), the first term on the right-hand side of (20) may be written

$$I(q_1,\ldots,q_n,q_{n+1};\text{---}) - (q_n+q_{n+1})\, I\left(\frac{q_n}{q_n+q_{n+1}},\frac{q_{n+1}}{q_n+q_{n+1}};\text{---}\right)$$

By (15), the last term on the right-hand side of (20) may be written

$$\begin{aligned} &(q_n+q_{n+1})\, I\left(\frac{q_n}{q_n+q_{n+1}},\frac{q_{n+1}}{q_n+q_{n+1}};\text{---}\right) + q_n I\left(\frac{p_{n1}}{q_n},\ldots,\frac{p_{nm}}{q_n};\text{---}\right) \\ &\quad + q_{n+1} I\left(\frac{p_{n+1,1}}{q_{n+1}},\ldots,\frac{p_{n+1,m}}{q_{n+1}};\text{---}\right) \end{aligned}$$

Plugging these two expressions into the appropriate places in (20), we obtain (19) for $n + 1$. Thus, (19) holds for every n.

Define[3]

$$f(r, r_0) \equiv I(1/r,\ldots,1/r,0,\ldots,0;1/r_0,\ldots,1/r_0) \tag{21}$$

By (19), with $m = r_0$ and $n = s_0$,

$$\begin{aligned} I(p_{11} = 1/rs,\ldots,p_{1r} = 1/rs, p_{1,r+1} = 0,\ldots,p_{1r_0} = 0,\text{---},p_{s1} = 1/rs,\ldots,p_{sr} = 1/rs, \\ p_{s,r+1} = 0,\ldots,p_{sr_0} = 0, p_{s+1,1} = 0,\ldots,p_{s_0r_0} = 0;\ 1/r_0s_0,\ldots,1/r_0s_0) \\ = I(1/s,\ldots,1/s,0,\ldots,0;1/s_0,\ldots,1/s_0) \\ + s(1/s)\, I(1/r,\ldots,1/r,0,\ldots,0;1/r_0,\ldots,1/r_0) \end{aligned}$$

[3] The symbols r, r_0, s, and s_0 will always mean positive integers, with $r_0 \geqslant r$ and $s_0 \geqslant s$.

Using (21) and (6), this becomes

$$f(rs, r_0s_0) = f(s, s_0) + f(r, r_0) \tag{22}$$

By (22) and (7),

$$f(rs, rs_0) = f(s, s_0) + f(r, r) = f(s, s_0) \tag{23}$$

Letting $r_0/r = r_0'/r_0' \equiv u$, (23) implies

$$f(r, r_0) = f(r, ur) = f(r, r_0'r/r') = f(r'r, r_0'r) = f(r', r_0') \tag{24}$$

Equation (24) says that $f(r, r_0)$ depends only on the ratio r_0/r. Thus,

$$g(r_0/r) \equiv f(r, r_0) \tag{25}$$

defines a single-valued function $g(x)$, defined for every rational $x \geqslant 1$. By (25) and (22), defining $x = r_0/r$ and $x' = r_0'/r'$,

$$g(xx') = g(r_0r_0'/rr') = f(rr', r_0r_0') = f(r, r_0) + f(r', r_0') = g(r_0/r) + g(r_0'/r')$$

Thus,

$$g(xx') = g(x) + g(x') \tag{26}$$

for every rational $x \geqslant 1$ and $x' \geqslant 1$. By (8), $g(x)$ is a monotonic increasing function.

We now show that

$$f(r, r_0) = g(r_0/r) = k \ln(r_0/r) \tag{27}$$

where k is a positive constant.

Proof of (27). By (26),

$$g(r_0^n) = ng(r_0) \tag{28}$$

Let integers n, r, and s be given arbitrarily, and let the integer m be determined by

$$m \leqslant n \ln s_0/\ln r_0 < m + 1 \tag{29}$$

from which

$$r_0^m \leqslant s_0^n < r_0^{m+1} \tag{30}$$

$$m/n \leqslant \ln s_0/\ln r_0 < (m + 1)/n \tag{31}$$

Since g is monotonic increasing, (28) and (30) imply

$$mg(r_0) \leqslant ng(s_0) \leqslant (m + 1)g(r_0)$$

so that

$$m/n \leqslant g(s_0)/g(r_0) \leqslant (m + 1)/n \tag{32}$$

By (31) and (32),

$$\left| \frac{g(s_0)}{g(r_0)} - \frac{\ln s_0}{\ln r_0} \right| \leqslant \frac{1}{n} \tag{33}$$

Since the left-hand side of (33) is independent of n, and since n can be chosen arbitrarily large, (33) implies that

$$g(s_0)/g(r_0) = \ln s_0/\ln r_0$$

or

$$g(r_0) = k \ln r_0 \tag{34}$$

Since g is monotonic increasing, the constant k must be positive. This proves (27) for $r = 1$ and any r_0. By using precisely the same reasoning as was used in obtaining (34), it may be seen that

$$g\left(\frac{r+1}{r}\right) = k \ln\left(\frac{r+1}{r}\right) \tag{35}$$

[Simply replace r_0 by $(r+1)/r$ and s_0 by $(s+1)/s$; Eqs. (28)–(34) all remain valid under this replacement.] We now prove (27) by induction on r. Assume (27) is valid for (r_0, r). Then, by (26) and (35),

$$g\left(\frac{r_0}{r}\right) = g\left(\frac{r_0}{r+1}\frac{r+1}{r}\right) = g\left(\frac{r_0}{r+1}\right) + g\left(\frac{r+1}{r}\right),$$

$$g\left(\frac{r_0}{r+1}\right) = g\left(\frac{r_0}{r}\right) - g\left(\frac{r+1}{r}\right) = k \ln\left(\frac{r_0}{r}\right) - k \ln\left(\frac{r+1}{r}\right) = k \ln\left(\frac{r_0}{r+1}\right)$$

Hence, if (27) holds for (r_0, r), then it holds for $(r_0, r+1)$. But, by (34), Eq. (27) holds for $(r_0, 1)$. Hence, (27) holds for all (r_0, r).

Let p and p^0 be rational numbers less than or equal to 1: $p = r/s \leqslant 1$, $p^0 = r_0/s_0 \leqslant 1$. If the three conditions $r_0 \geqslant r$, $s_0 \geqslant s$, and $s_0 - r_0 \geqslant s - r$ do not hold for the choice (r, s, r_0, s_0), then replace (r_0, s_0) by (nr_0, ns_0), with n sufficiently large that the three conditions hold for (r, s, nr_0, ns_0). This replacement does not affect p or p^0. Hence, it may be assumed, without loss of generality, that $r_0 \geqslant r$, $s_0 \geqslant s$, $s_0 - r_0 \geqslant s - r$. By (6) and (19), with appropriate grouping of the variables,

$$\begin{aligned} &I(1/s,\ldots, 1/s, 0,\ldots, 0; 1/s_0,\ldots, 1/s_0) \\ &\quad = I[r/s, (s-r)/s; r_0/s_0, (s_0 - r_0)/s_0] \\ &\qquad + (r/s)\, I(1/r,\ldots, 1/r, 0,\ldots, 0; 1/r_0,\ldots, 1/r_0) \\ &\qquad + [(s-r)/s]\, I[1/(s-r),\ldots, 1/(s-r), 0,\ldots, 0; 1/(s_0 - r_0),\ldots, 1/(s_0 - r_0)] \end{aligned}$$

By (21) and (27), this becomes

$$\begin{aligned} I(p, 1-p; p^0, 1-p^0) &= f(s, s_0) - pf(r, r_0) - (1-p) f(s-r, s_0 - r_0) \\ &= kp \ln(p/p^0) + k(1-p) \ln[(1-p)/(1-p^0)] \end{aligned}$$

By (5), this result extends to all irrational p. This proves the theorem for $n = 2$. Using (16), the theorem follows by induction on n.

ACKNOWLEDGMENTS

The author would like to thank the Aspen Center for Physics for providing a stimulating environment during the completion of this work, and the National Science Foundation for providing research support.

REFERENCES

1. C. E. Shannon, *Bell System Tech. J.* **27**, 379 and 623 (1948). Reprinted in C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1949).
2. A. Feinstein, *Foundations of Information Theory* (McGraw-Hill Book Company, Inc., New York, 1958).
3. A. I. Khinchin, *Mathematical Foundations of Information Theory* (Dover Publications, Inc., New York, 1957).
4. F. M. Reza, *An Introduction to Information Theory* (McGraw-Hill Book Company, Inc., New York, 1961).
5. A. Katz, *Principles of Statistical Mechanics: The Information Theory Approach* (W. H. Freeman and Company, San Francisco, 1967).
6. S. Kullback, *Information Theory and Statistics* (John Wiley & Sons, Inc., New York, 1959).
7. A. Renyi, *Wahrscheinlichkeitrechnung* (VEB Deutscher Verlag der Wissenschaften, Berlin, 1966).
8. M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, translated and edited by A. Feinstein (Holden-Day, Inc., San Francisco, 1964).
9. F. Schlögl, in *Statistical Mechanics* (Proc. of the IUPAP Meeting, Copenhagen, 1966), edited by T. A. Bak (W. A. Benjamin, Inc., New York, 1967); in *Proc. of the International Conf. on Statistical Mechanics* (Kyoto, 1968) (supplement to *J. Phys. Soc. Japan* **26**, 1969).
10. A. Hobson, *Concepts in Statistical Mechanics* (Gordon and Breach, Science Publishers, Inc., 1970).