Aequationes Mathematicae 27 (1984) 1–19
University of Waterloo

0001-9054/84/001001-19$1.50 + 0.20/0

Survey Paper

Measuring information beyond communication theory—Why some generalized information measures may be useful, others not

J. ACZÉL

Summary (and Keywords). Non-communication models for information theory: games and experiments. Measures of uncertainty and information: entropies, divergences, information improvements.

Some useful properties of information measures, symmetry, bounds, behaviour under composition, branching, conditional measures, sources. Rényi measures, measures of higher degree.

Promising and not so promising generalizations. Measures which depend not just upon the probabilities but (also) upon the subject matters.

1. The usual model for information theory in communication theory consists of messages, their frequencies (or frequencies of errors in their transmission), coding, and the entropy as lower bound of the average codeword length. While this is where information theory came from, as it happens, this is not a very universal model and its use beyond information theory is limited. Questionnaire theory is rather close to this model but I will not deal with that theory here.

A connection exists with a very flexible model, that of logical games like "twenty questions" and "counterfeit coins". Indeed the "twenty (or n) questions" or "binary search" can be made into a highly effective tool, similar to urn model in probability theory. The yes (*1*) or no (*0*) answers may be considered as coding symbols and the string of *1*'s (yes) and *0*'s (no) as determining the searched object by its codeword. Similarly, whether the left (*2*) or right (*1*) beam of the beam-balance used for finding a counterfeit coin is lighter or they are equal (*0*) furnishes the symbols, and their sequence the codeword of that coin.

I think, however, that the most convenient model for our purposes is that of an experiment with possible outcomes (events) $E_1, E_2, \ldots, E_n$ which have probabilities $p_1, p_2, \ldots, p_n (\sum_{k=1}^{n} p_k = 1)$. Traditionally the Shannon entropy [40, 9]

$$H_n(p_1, \ldots, p_n) = -\sum_{k=1}^{n} p_k \log p_k \tag{1}$$

AMS (1980) subject classification: Primary 94A15, 94A17, Secondary 39B40, 39B70.

Manuscript received May 3, 1983, and in final form, August 4, 1983.

(usually $\log = \log_2$) measures the amount of uncertainty in the outcome of the experiment—and also of the information expected from this experience (for short, uncertainty and expected information).

2. This interpretation is explained by properties of the expression (1) [9]. For instance (1) is *maximal* ($=\log n$) when $p_1 = \ldots = p_n \left(= \frac{1}{n}\right)$, that is, when the probabilities are equal, as can be expected of a measure of uncertainty. (The *redundancy of first order* is thus nonnegative: $1 - \frac{H_n}{\log n} \geqq 0$.) More trivially, H_n is *nonnegative* (thus *bounded* on both sides). It is also *symmetric* (it does not depend upon the order of labelling the events).

More interesting are the properties describing the behaviour of this measure *under composition*, that is, when two (or more) experiments are performed:

$$H_{mn}[p(E_1, F_1), \ldots, p(E_1, F_n), \ldots, p(E_m, F_1), \ldots, p(E_m, F_n)] \leqq$$
$$H_m[p(E_1), \ldots, p(E_m)] + H_n[p(F_1), \ldots, p(F_n)]$$

which is called *subadditivity with equality* (*additivity*) when the two experiments are *independent*, that is when $p(E_i, F_j) = p(E_i)p(F_j)$ ($i = 1, \ldots, m$; $j = 1, \ldots, n$). Here $p(E_i)$, $p(F_j)$ and $p(E_i, F_j)$ are the probabilities of the outcome of E_i in the first experiment, of F_j in the second, or of the outcome E_i in the first *and* F_j in the second experiment, respectively. The subadditivity and additivity can be written shorter as

$$H(PQ) \leqq H(P) + H(Q) \tag{2}$$

and

$$H(PQ) = H(P) + H(Q) \text{ when } P \text{ and } Q \text{ are independent.} \tag{3}$$

They state that the information expected from two experiments is not more than the sum of the amounts of information expected from the individual experiments and equal to this sum if the two experiments are independent.

In analogy to the definition of the conditional probability $p(F/E)$ by $p(E, F) = p(E)p(F/E)$, the *conditional entropy* $H(Q/P)$ can be defined by

$$H(PQ) = H(P) + H(Q/P). \tag{4}$$

In view of this, (2) and (3) can be written as

$$H(Q/P) \leqq H(Q) \tag{5}$$

and

$$H(Q/P) = H(Q) \text{ if } P \text{ and } Q \text{ are independent,} \tag{6}$$

respectively, again conforming to our intuitive expectations of what *conditional uncertainty* should mean. Thus the mutual information $I(P, Q) = H(Q) - H(Q/P) = H(P) + H(Q) - H(PQ)$ is nonnegative (0 if P and Q are independent). The *channel capacity* is the maximum of the mutual information between the input and the output (on all possible inputs considered).

For Shannon entropies, the relatively simple formula

$$H(Q/P) = \sum_{i=1}^{m} p(E_i) H_n[p(F_1/E_i), \ldots, p(F_n/E_i)] \tag{7}$$

holds. Also, for these conditional entropies we have the following generalization of (2):

$$H(PQ/R) \leqq H(P/R) + H(Q/R). \tag{8}$$

(In view of (6), this reduces to (2) if R is independent of P and Q.) By (4) we can write (8) as

$$H(PQR) + H(R) \leqq H(PR) + H(QR). \tag{9}$$

Again by (4), this can also be written as

$$H(P/QR) \leqq H(P/R), \tag{10}$$

which is a generalization of (5). For Shannon entropies we also have

$$H(P/P_1 P_2 \ldots P_r) \leqq H(P/P_1 P_2 \ldots P_{r-1})$$

as a further generalization of (5) and (10).—The conditional entropies are also nonnegative. These equations and inequalities make the definition of the *source entropy* as

$$H^{(\infty)} = \lim_{r \to \infty} H(P^r)/r = \lim_{r \to \infty} H(P/P^{r-1})$$

possible (and the existence and equality of the two limits valid). For our model this

means that the *average information expected from r repetitions* of an experiment and the *conditional uncertainty in the result of the r-th repetition* have limits and they are equal. These conditional uncertainties decrease with increasing r and are not greater than the unconditional uncertainty in the original experiment and thus than $\log n$. So the *redundancy* (of infinite order) is also nonnegative:

$$1 - \frac{H^{(\infty)}}{\log n} \geqq 0.$$

The definition (1) of the Shannon entropy makes no sense if zero probabilities ($p_k = 0$ for one or more k, but not for all k, because $\sum_1^n p_k = 1$) are permissible. But it can be made meaningful in these cases too, if the convention

$$0 \log 0 = 0 \tag{11}$$

is adopted. Then the *expansibility*

$$H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n) \tag{12}$$

holds too, stating that the uncertainty does not change when we add to the possible outcomes of the experiment one with probability 0. This sounds pretty intuitive too, though not for all applications (for instance not in economics, where (1) is interpreted as measure of equality, p_k being there the relative wealth or income, etc. of the k-th group, cf. [43]). Anyway, the expansibility (12) and (4) with (7) imply

$$H_{n+1}(p_1 q_1, p_1 q_2, p_2, \ldots, p_n) = H_n(p_1, p_2, \ldots, p_n) + p_1 H_2(q_1, q_2) \tag{13}$$

($\sum_1^n p_k = 1 = q_1 + q_2$). This is the *recursivity* (or *branching* property; the latter name, however, has been used recently for somewhat more general properties, see below), because it completely determines H_n recursively for all $n \geqq 2$ if H_2 is known. It describes what happens to the uncertainty if one possible outcome of the experiment is split into two and so makes intuitive sense also without (12), (7) and (4).

3. Already a few of these properties are sufficient to *characterize* the Shannon entropy. For instance [11, 9], only

$$a \log \#(p_k \neq 0) + b \sum_1^n p_k \log p_k \quad (a \geqq 0 \geqq b)$$

is symmetric, subadditive (2), additive (3) and expansible (12). In the second term the convention (11) is used and it is, of course, a constant multiple of the Shannon entropy. The first term is (a constant multiple of) the so called Hartley entropy, the logarithm of the number of nonzero probabilities. However if, instead of (2), we suppose (8) or, equivalently, (9) [or (10)] then, as B. Forte has recently observed, $a = 0$ and we have characterized the Shannon entropies (up to a nonnegative multiplicative constant; we can get rid of it if we use the normalizing condition $H_2(\frac{1}{2}, \frac{1}{2}) = 1$, for example).

On the other hand [20, 23, 35], constant multiples of the Shannon entropy are the only bounded symmetric and recursive (13) entropies. Contrary to the first, this characterization works [13] also if zero probabilities are excluded, which may be desirable in view of our above observation on applications in economics (and of the generalizations to follow). The boundedness condition here may be supposed only for H_2, only on a (small) interval (or even only on a set of positive measure, but his, in my opinion, is only of theoretical interest) or replaced [9] by

$$\lim_{q \to 0+} H_2(1-q, q) = 0 \tag{14}$$

This last property is intuitive again: It states that, for experiments with two possible results, if one is very probable, the other very improbable, then the uncertainty in the outcome of the experiments is small. The properties (14), (2), (3), (12) and symmetry give another characterization of nonnegative multiples of the Shannon entropy.

A further characterization comes from forecasting theory. The probabilities $p_1, \ldots, p_n$ of the events $E_1, \ldots, E_n$ (the objects of the forecasts, weather, market situations — or the outcomes of an experiment) are estimated by a forecaster as $q_1, \ldots, q_n$. A payoff $f(q_k)$ is paid to the forecaster if E_k happens. The question "how to keep the forecaster honest?" is answered by choosing the payoff function so that the forecaster's expected gain is maximal when $q_k = p_k$, that is,

$$\sum^{n} p_k f(q_k) \leqq \sum^{n} p_k f(p_k). \tag{15}$$

If (15) is supposed for all n or even for one $n > 2$, then [24, 9] $f(q) = a \log q + b$ ($a \geqq 0$) is obtained without regularity conditions on f and without 0-probabilities, so the maximal expected gain is

$$\sum^{n} p_k f(p_k) = a \sum_{k=1}^{n} p_k \log p_k + b,$$

containing the Shannon entropy. Conversely, the $f(q)$'s just obtained satisfy (15). This is a consequence of the well known nonnegativity of the directed divergence,

$$\sum^{n} p_k \log \frac{p_k}{q_k} \tag{16}$$

($\sum^{n} p_k = \sum^{n} q_k = 1$), another important measure of information (this containing two rather than one probability distribution, it has applications a.o. in statistics), which is thus characterized by (15) [9]. Obvious further generalizations of (16) have been abundantly treated. It seems that Theil's [43] information improvement (containing three probability distributions and introduced for the purpose of applications in economics)

$$\sum^{n} p_k \log \frac{r_k}{q_k} \tag{17}$$

($\sum^{n} p_k = \sum^{n} q_k = \sum^{n} r_k = 1$) is the last where we can hope for genuine applications.

These measures have also been characterized. Here the problem of some $q_k = 0$ (or $r_k = 0$) is not so easily eliminated as in (1) by (11). So rather complicated prescriptions were attached, also to the characterizations, like the requirement that $q_{k_0} = 0$ (or $r_{k_0} = 0$) should imply $p_{k_0} = 0$. (Several "characterizations" where these restrictions were missing, were simply incorrect.) Fortunately, again, a characterization by an analogue of the recursivity (13) with several probability distributions, all containing positive probabilities only, has been found recently [13] which frees us from such restrictions. Actually one characterizes e.g. $\sum^{n} p_k (a \log p_k + b \log q_k + c \log r_k)$ and then specializes it to, say, (17) by normalizing conditions.

4. What is the use of such characterization theorems? They list the properties (in a "good" theorem this list is short and the properties are simple and intuitive) which have to be checked for a quantity which occurs *in a new field of applications* in order to notice that it is really the Shannon entropy (or a similar measure depending upon more than one probability distributions).

On the other hand, other, *more flexible measures* may be needed in these new fields (or in the old ones). Since the above conditions do characterize the Shannon entropy, other entropies cannot have *all* these properties. Some have to be dropped or modified. The recursivity (13) rather suggests the generalization

$$H_{n+1}(p_1 q_1, p_1 q_2, p_2, \ldots, p_n) = H_n(p_1, \ldots, p_n) + m(p_1) H_2(q_1, q_2) \quad (n = 2, 3, \ldots). \tag{18}$$

How should we choose the function m? If the similar equation

$$H_{n+2}(p_1q_1, p_1q_2, p_1q_3 \cdot p_2, \ldots, p_n) = H_n(p_1, \ldots, p_n) + m(p_1)H_3(q_1, q_2, q_3) \; (n = 2, 3, \ldots)$$

is also supposed, then, as can be easily proved, m has to be *multiplicative* ($m(pq) = m(p)m(q)$). In this case (18) is completely solved, again also for several probability distributions containing only positive probabilities [13].

Practically, only the cases where m is not identically 0 and somewhat regular (say, measurable) seem to be applicable. Then $m(p) = p^a$ and our measures are *of degree a.* In this case, if $a \neq 1$, the general symmetric solution of (18) is, interestingly without any boundedness or other regularity suppositions,

$$H_n^a(p_1, p_2, \ldots, p_n) = c\left(\sum p_k^a - 1\right) \tag{19}$$

the *entropy of degree a* [29, 18, 9]. If there are several (s) probability distributions, then the p_k are (s-dimensional) vectors (with $\sum_{k=1}^{n} p_{kj} = 1$ for $j = 1, \ldots, s$) and in (19)

$$p_k^a = (p_{k1}, \ldots, p_{ks})^{(a^1, \ldots, a_j)} = \prod_{j=1}^{s} p_{kj}^{a_j}.$$

In the one-dimensional (one probability distribution) case (and also in some components of more dimensional vectors) the result (19) can be extended by the convention $0^a = 0$ (even if $a \leqq 0$). — If $s = 1$, $a > 1$ (and $c \leqq 0$ so that $H_n \geqq 0$), then the entropies of degree a are *subadditive* (2), but in general not additive (3).—If we normalize (19) (just as the Shannon entropy), by $H_2^a(\frac{1}{2}, \frac{1}{2}) = 1$ then $c = (2^{1-a} - 1)^{-1}$ and ($s = 1$)

$$H_n^a(p_1, \ldots, p_n) = (2^{1-a} - 1)^{-1}\left(\sum p_k^a - 1\right) \tag{20}$$

and the limit of this, as $a \to 1$, is the Shannon entropy (1) (which, as we have seen, is subadditive too but also additive). This limit relation, the subadditivity of H^a for $a > 1$ and applications [29, 21, 22], for instance to pattern recognition, seem to indicate that entropies of degree $a > 1$ (and other, similar, multidimensional information measures) could have fruitful applications elsewhere too. The lack of additivity may be compensated by the flexibility assured by the possibility of choosing the parameter a according to the special characteristics of the specific problem.

5. There are other characterizations of the Shannon entropy and of similar

measures based on the formal *shape* of the expressions (1), (16), (17), etc. combined with *additivity* (or with a generalized additivity).

For instance, both (1) and (19) (written as $\sum^{n} c(p_k^a - p_k)$) are of the *sum-form*

$$H_n(p_1, \ldots, p_n) = \sum^{n} \phi(p_k). \tag{21}$$

This may seem to be a rather superficial property, but a consequence of [37] (cf. [9]) shows that (21) follows from the symmetry, expansibility (12) of H_n and from the "ultimate generalization" of (13) and (18), the *branching property*

$$H_n(p_1, p_2, p_3, \ldots, p_n) = H_{n-1}(p_1 + p_2, p_3, \ldots, p_n) + J_n(p_1, p_2),$$

or, if no expansibility is supposed (for instance when 0 probabilities are excluded), from the "penultimate generalization"

$$H_n(p_1, p_2, p_3, \ldots, p_n) = H_{n-1}(p_1 + p_2, p_3, \ldots, p_n) + J(p_1, p_2) \quad (n = 3, 4, \ldots).$$

(In order to see that these are indeed generalizations of (18) and thus also of (13), write (18) in the form

$$H_{n+1}(\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \ldots, p_{n+1}) = H_n(\tilde{p}_1 + \tilde{p}_2, \tilde{p}_3, \ldots, \tilde{p}_{n+1})$$
$$+ m(\tilde{p}_1 + \tilde{p}_2) H_2\left(\frac{\tilde{p}_1}{\tilde{p}_1 + \tilde{p}_2}, \frac{\tilde{p}_2}{\tilde{p}_1 + \tilde{p}_2}\right),$$

where $\tilde{p}_1 = p_1 q_1$, $\tilde{p}_2 = p_1 \tilde{q}_2$, and $\tilde{p}_j = p_{j-1}$ for $j = 3, 4, \ldots, n+1$).

If H_n is additive and of the form (21) with some mildly regular (measurable) ϕ, then it is a constant multiple of the Shannon entropy (1) [19, 9]. (Even additivity can be weakened by supposing it only for particular m, n.)

How does (19) fit into all this? It is of the form (21) but not additive. Instead, it satisfies

$$H(PQ) = H(P) + H(Q) + \frac{1}{c} H(P)H(Q) \text{ if } P \text{ and } Q \text{ are independent.} \tag{22}$$

This and (21) with measurable ϕ again characterize the entropies (19). There are also similar theorems for measures depending upon several probability distributions (see e.g. [30]). The property (22) is simple enough and reduces to (3) if $c \to \infty$ (in accordance with $c = (2^{1-a} - 1)^{-1} \to \infty$ as $a \to 1$) but I see no natural interpretation. New measures

of information, based on further generalizations of (21) and (22) seem even less natural and/or applicable to me.

6. Another way of looking at the Shannon entropy (1) is to notice that it is the *arithmetic mean* of the $(-\log p_k)$, *weighted by the probabilities* p_k (notice that $\sum^n p_k = 1$). good case was made by Rényi [38] that the $(-\log p_k)$ here is the *information* yielded by the event E_k with probability p_k (here again it is of advantage to exclude 0 probabilities), which indeed makes (1) the *expected information.* It was further proposed in [38, 39] to consider also other *quasiarithmetic means* of these $(-\log p_k)$, thus defining

$$^{\psi}H_n(p_1, \ldots, p_n) = \psi^{-1}\left(\sum^n p_k \psi(-\log p_k)\right)$$

where $\psi:]0, \infty[\to \mathbb{R}$ is continuous and strictly increasing. If $^{\psi}H$ is of this form, additive, and satisfies $\lim_{q \to 0+} {}^{\psi}H_2(1-q, q) = 0$ (cf. (14)), then [17, 7, 9] either $^{\psi}H_n$ is the Shannon entropy or the Rényi entropy of order $a > 0$ $(a \neq 1)$

$$_aH_n(p_1, \ldots, p_n) = \frac{1}{1-a} \log \sum^n p_k^a. \tag{23}$$

In fact, $_aH_n$ is additive even for negative a. Here too, both the entropy and its characterization can be extended to 0-probabilities. Also, just as for (20), the limit of (23) as $a \to 1$ is again the Shannon entropy. Moreover, the limit of (23) as $a \to 0$ is the Hartley entropy $\log \#(p_k \neq 0)$. These limit relations, the additivity, applications to "random search" problems [39] and rather similar relations to other mean codeword lengths as those of the Shannon entropy to the arithmetic mean codeword length [14, 15, 1, 9] make also the Rényi entropies good candidates for a wide range of applications.

Again, also measures of order a, depending upon more than one probability distribution can be defined. For instance, the analogue of (16) is the *directed divergence of order a*

$$\frac{1}{a-1} \log \sum^n p_k^a q_k^{1-a} \tag{24}$$

which converges to (16) as $a \to 1$. For a characterization of (24) and (16), similar to that of $_aH_n$ above, see [38].

There is also a connection between (24) and (15) [25, 9]: For positive f,

$$\sum^{n} p_k \frac{f(p_k)}{f(q_k)} \leqq 1 \tag{25}$$

if, and only if, $f(q) = cq^{a-1}$ with $c > 0$, $0 \leqq a \leqq 1$. This f indeed satisfies (25): trivially

$$\left(\sum^{n} q_k = 1 = \sum^{n} p_k\right) \text{ for } a = 0,\ 1 \text{ and}$$

$$\sum^{n} p_k \frac{p_k^{a-1}}{q_k^{a-1}} \leqq 1 \text{ if } 0 < a < 1$$

by Hölder's inequality. This exactly expresses that the *directed divergence of order a* (24) *is nonnegative* for $0 \leqq a < 1$. *This is true also for* $a > 1$ because, again by the Hölder inequality,

$$\sum^{n} p_k \frac{p_k^{a-1}}{q_k^{a-1}} \geqq 1 \text{ if } a > 1.$$

See [26] for the corresponding counterpart of the inequality (25), viz.

$$\sum^{n} p_k \frac{f(p_k)}{f(q_k)} \geqq 1.$$

In all these results zero probabilities can and should be excluded.

7. There is a large number of "entropies" and other "information measures" and their "characterizations", mostly formal generalizations of (1), (19), (16), (24), (17), (23) etc. popping up almost daily in the literature. It may be reassuring to know that most are and will in all probability be completely useless. Just possibly the following (and the respective measures depending upon more than one probability distribution) may be exceptions. The entropies of order (a, b) [8, 33]:

$$\frac{1}{b-a} log\left(\sum^{n} p_k^a / \sum^{n} p_k^b\right) \quad (b \neq a), \qquad -\sum^{n} p_k^a \log p_k / \sum^{n} p_k^a \quad (\text{for } b = a),$$

those of degree (a, b) [42]:

$$(2^{1-a} - 2^{1-b})^{-1} \sum^{n} (p_k^a - p_k^b) \quad (b \neq a), \qquad -2^{a-1} \sum^{n} p_k^a \log p_k \quad (\text{for } b = a)$$

and those of order a and rank b (see [41, 44], also for "characterizations" by (22) and by a generalized quasiarithmeticity $H(p_1, \dots p_n) = \psi^{-1}[\sum^n \Phi(p_k)\psi(-\log p_k)/\sum^n \Phi(p_k)]$):

$$H(p_1, \dots, p_n; a, b) = (2^{(1-a)b} - 1)^{-1}((\sum^n p_k^a)^b - 1) \quad (a \neq 1, b \neq 0).$$

The first two seem interesting because they are so natural generalizations of (23) ${}_aH = \frac{1}{1-a}\log(\sum^n p_k^a / \sum^n p_k)$ or of (20) $H^a = (2^{1-a} - 1)^{-1}\sum^n (p_k^a - p_k)$, respectively, the third because it contains as special or limiting cases both the entropies of degree *a and* of order a (just *one* such relation would *not* justify the introduction of a new measure):

$$H(p_1, \dots, p_n; a, 1) = H^a(p_1, \dots, p_n) \text{ and } \lim_{b \to 0} H(p_1, \dots, p_n; a, b) = {}_aH(p_1, \dots, p_n).$$

The Shannon entropy is also directly a limiting case (not just through ${}_aH$ and H^a):

$$\lim_{a \to 1} H(p_1, \dots, p_n; a, b) = -\sum^n p_k \log p_k.$$

I wish to urge here caution with regard to generalizations in general, and in particular with regard to those introduced through characterizations. In the best of all possible worlds, there is an information measure, which originated from an applied problem, it has interesting properties (usually attractive, *reasonable* generalizations of properties of Shannon's entropy or of similar widely used measures), and these properties characterize it. Less ideal, but still acceptable in my opinion, is the following situation. Some natural looking weakening or generalization of the properties characterizing Shannon-type measures are isolated and all measures having these properties are determined. If the properties are indeed intuitive and significant then there is a good chance that the measures thus obtained may have future applications. But what many authors seem to do is to contrive some generalization of known information measures (usually by sticking in more parameters almost at random here or there), derive its often not very interesting or natural and also often not very attractive properties and then characterize, by several of these properties, the "measures" which they have defined in the first place. Not many good or useful results can be expected from this kind of activity.

A generalization of (16), which does seem fruitful is the f-divergence $\sum^n q_k f\left(\frac{p_k}{q_k}\right)$ introduced and characterized by Csiszár [16], where f is an arbitrary convex function (in (16), $f(x) = x \log x$).

8. There are, however, generalizations in a completely different vein. One is the "theory of information without probability" (see e.g. [27]). It is based on the observation that some events furnish information even though they have no probabilities, because they cannot be repeated. Part of this is connected to the outdated or at least partisan view that probabilities are "limits" of frequencies (in a certain sense, difficult to define since late large variations are always possible), so only such events can have probabilities which can be repeated infinitely often (or at least as often as we want to). But this would be analogous to calling only such lengths, temperatures, etc. measurable which can be measured arbitrarily often.

Since Kolmogorov's fundamental work [34] in the 1930's, it is more or less accepted that probability is what satisfies certain conditions (axioms). According to our above definition of $I = -\log p$ as the information yields by the event E, if it has a probability p, we can conversely define 2^{-I} as the (generalized) probability of E if I is given. But the main objection against that theory seems to be that it does not lead far enough (which, on the other hand, could be construed also as an incentive to work on it harder). We will not deal here further with that subject, rather we give some details on another generalization of the probabilistic theory of information, the "mixed theory".

In the *mixed theory of information*, the measures of information are permitted to depend both on the events (messages, outcomes of an experiment, weather, market situations, etc.) themselves and on their probabilities (or similar parameters). For this we have to grasp mathematically the concept of an "event". It seems that they can be considered adequately as elements of a *ring of sets*, really of subsets of a comprehensive set (universe) S which contains, with any two subsets, also their union ($\cup$) and difference ($\setminus$), therefore also their intersection ($\cap$) and the $\emptyset$ (empty) set. If also S belongs to the ring then it is *Boolean*. So we are dealing now with (real valued) measures

$$H_n\begin{Bmatrix} E_1, E_2, \dots, E_n \\ p_1, p_2, \dots, p_n \end{Bmatrix} \quad (E_j \cap E_k = \emptyset \text{ if } j \neq k,\ p_k \geqq 0,\ \sum^n p_k = 1).$$

(Again, we may have also further sets of probabilities but those should be positive, at least when the corresponding p is.) By *symmetry* we mean here that the value of H_n does not change if two *columns* $\begin{pmatrix} E_j \\ p_j \end{pmatrix}$ and $\begin{pmatrix} E_k \\ p_k \end{pmatrix}$ are exchanged. The analogue of the recursivity (13) is here

$$\left.\begin{aligned} &H_{n+1}\begin{Bmatrix} E_1 \cap F_1, E_1 \cap F_2, E_2, \dots, E_n \\ p_1 q_1, p_1 q_2, \quad p_2, \dots, p_n \end{Bmatrix} = H_n\begin{Bmatrix} E_1, E_2, \dots, E_n \\ p_1, p_2, \dots, p_n \end{Bmatrix} + p_1 H_2\begin{Bmatrix} F_1, F_2 \\ q_1, q_2 \end{Bmatrix} \\ &(E_j \cap E_k = \emptyset \text{ if } j \neq k,\ F_1 \cap F_2 \neq 0,\ p_k \geqq 0,\ q_1 \geqq 0,\ q_2 \geqq 0,\ \sum^n p_k = 1 = q_1 + q_2). \end{aligned}\right\} \tag{26}$$

Again one may ask for the general form of symmetric measures satisfying (26) and, say, continuous in the probabilities. This has been proved [10, 5] to be

$$a\sum^{n} p_k \log p_k + \sum^{n} p_k g(E_k) - g\left(\bigcup^{n} E_k\right) \tag{27}$$

where a is an arbitrary constant and g an arbitrary real valued function on the ring of sets. Note that it is not supposed here that $\bigcup^{n} E_k = S$ (S may not even be in the ring of sets considered). If we have Boolean rings (thus containing S) and $\bigcup^{n} E_k = S$ then, with $h(E) = g(E) - g(S)$, (27) goes over into

$$a\sum^{n} p_k \log p_k + \sum^{n} p_k h(E_k). \tag{28}$$

(For the characterization of similar quantities depending upon several probability distributions, see [31, 32].)

The quantities (27) and (28), consisting of Shannon's entropy plus additional terms depending upon events, are called *inset entropies* of the Shannon type. (Inset may be understood in its dictionary meaning "a map set into a map" or as "in set", but really the name was chosen because the idea was born at a meeting at the École Normale Supérieure de l'Enseignment Technique — ENSET — near Paris.)

9. There are applications of these measures in information theory and elsewhere. Traditionally

$$-\int_a^b f(x)\log f(x)dx \tag{29}$$

is considered to be the "continuous analogue" of (1) for a random variable with probability density function f. Contrary to what one may think, however, the sums approximating (29)

$$-\sum^{n} f(\xi_k)\log f(\xi_k)(x_k - x_{k-1}) \tag{30}$$

($a = x_0 \leqq \xi_1 \leqq x_1 \leqq \xi_2 \leqq \ldots \leqq \xi_n \leqq x_n = b$) are *not* Shannon entropies of the form (1) but they *are* inset entropies of the form (28). We see this by realizing that here

$$p_k = f(\xi_k)(x_k - x_{k-1}),\ E_k =]x_{k-1}, x_k],\ \left(\bigcup_{k=1}^{n} E_k =]a, b]\right) \text{ and } x_k - x_{k-1} = l(E_k)$$

(l standing for length) so that (30) goes over into

$$-\sum^{n} p_k \log p_k + \sum^{n} p_k \log l(E_k) \tag{31}$$

which is indeed of the form (28) [4].

However, contrary to (1), the quantity (29) and therefore also (31) may be negative. (That is one reason why (29) is not a very good analogue of the Shannon entropy (1) and also why in characterizations of the above inset entropies we have used continuity rather than boundedness—we certainly could not have used nonnegativity.) On the other hand, if $n = 1$ and $E =]a, b]$, then $p_1 = 1$ and (31) reduces to $\log l(E)$. This corresponds to the case where we know that the value of the random variable falls into $]a, b]$ but don't know its probability distribution. Since $]a, b] = \bigcup^{n}]x_{k-1}, x_k] = \bigcup_{k=1}^{n} E_k$, the decrease of uncertainty between this ignorant state described by (31), where we know the probabilities, is

$$\log l\left(\bigcup^{n} E_k\right) - \sum^{n} p_k \log l(E_k) + \sum^{n} p_k \log p_k, \tag{32}$$

an inset entropy of the form (27) (this happens to be also nonnegative) [4].

Also the $q_k = l(E_k)/l\left(\bigcup^{n} E_k\right)$ can, of course, be considered (geometric) probabilities, so (32) may be written as $\sum p_k \log \frac{p_k}{q_k}$, the quantity (16), which we have encountered before and called directed divergence. Conversely, since in inset entropies the dependence upon the events can appear as dependence upon *parameters* determined by the events, *all* directed divergences, information improvements, etc. may be considered inset entropies.

There have been recent efforts to take, in addition to the probabilities, also the "usefulness" of events (again of weather, of market situations, etc.) into consideration [28]. In view of the above it would seem worthwhile to look at these "measures of useful information" as inset measures and thus develop an appropriate theory.

For a more playful application, to the theory of gambling, Meginnis [36] considers the *second* term in (28) the expected gain (E_k being the *k-th* outcome of the game which has a gain in the amount $h(E_k)$ attached to it). So this time it is the *first* term which has to be explained. Since the expected gain alone would not motivate gambling (it is almost always nonpositive), he interprets the first term as quantifying the *joy in gambling*. He too derives (28) from requirements fairly natural for this application.

Also *inset measure of degree a* can be characterized by analogues (of symmetry

and) of the recursivity (18) with $m(p) = p^a$ ([12, 32] in the one dimensional and W. Sander in [45], p. 282, in the multidimensional case). For inset entropies of degree a this gives the expression

$$\sum^n p_k^a g(E_k) - g\left(\bigcup^n E_k\right)$$

or, if $\bigcup^n E_k = S$,

$$\sum^n p_k^a h(E_k) - c,$$

(g, h arbitrary real valued functions on the ring of sets, c an arbitrary constant), which are even simpler than (27) and (28). Also the latter has applications to the theory of gambling [36].

10. Finally, also the forecasting theory application, mentioned above, can be naturally generalized to such inset situations. The payoff $f(E_k, q_k)$ may very well depend upon the *k-th* event itself, not only on its probability. Then the forecaster's expected gain is

$$\sum^n p_k f(E_k, q_k)$$

and we "keep the forecaster honest" by choosing f so that (cf. (15))

$$\sum^n p_k f(E_k, q_k) \leqq \sum^n p_k f(E_k, p_k). \tag{33}$$

It has been proved [3] that this happens (for a fixed $n > 2$) if, and only if,

$$f(E, q) = a \log q + h(E)$$

($a \geqq 0$ an arbitrary constant, h an arbitrary real valued function on the ring of events). So the right hand side of (33) goes over into

$$a \sum^n p_k \log p_k + \sum^n p_k h(E_k),$$

again an inset entropy of the form (28).

Not all characterizations of the Shannon and other probabilistic measures can be easily extended to characterize some inset measures. For instance, the obvious generalization of the sum property (21) to $H(P) = \sum^{n} \phi(E_k, p_k)$ with ϕ measurable in the probabilities and with the additivity similarly generalized from (3) implies that ϕ does not depend upon the E_k and gives essentially only multiples of the purely probabilistic Shannon entropy [2].

The characterization theory of inset measures of information is also somewhat behind the probabilistic theory. Among others, as we have mentioned before, 0-probabilities have recently been eliminated from much of the characterization theory of purely probabilistic information measures. But in the similar theory of inset measures, impossible events (empty sets) are vigorously used. While the probabilities are separate variables, it seems to be common sense to require that impossible events have 0 probabilities. Up to now there are only two results where impossible events and 0 probabilities have been completely eliminated. One is the determination by B. Ebanks and Gy. Maksa in [45] (pp. 269 and 277) of all one dimensional measures (entropies) which satisfy the inset analogue of (18) with $m(p) = p^a$, *that is*,

$$H_{n+1}\begin{Bmatrix} E_1 \cap F_1, E_1 \cap F_2, E_2, \ldots, E_n \\ p_1 q_1, p_1 q_2, p_2, \ldots, p_n \end{Bmatrix} = H_n\begin{Bmatrix} E_1, E_2, \ldots, E_n \\ p_1, p_2, \ldots, p_n \end{Bmatrix} + p_1^a H_2\begin{Bmatrix} F_1, F_2 \\ q_1, q_2 \end{Bmatrix}$$

$$(E_k \neq \emptyset,\ F_1 \neq \emptyset,\ F_2 \neq \emptyset, E_j \cap E_k \neq \emptyset, \text{ if } j \neq k,\ F_1 \cap F_2 = \emptyset,\ p_k > 0,\ q_1 > 0,\ q_2 > 0,$$

$$\sum^{n} p_k = 1 = q_1 + q_2)$$

(cf. (26)). The general case (arbitrary multiplicative function m, arbitrary dimensions, is still not solved. The other result is the solution of (33) in [3], mentioned above, keeping the expert honest with inset reward and without impossible events and zero probabilities (but still with some money). Some simple results in [6] have already been used and may serve as tools to obtain other inset characterization theorems without empty sets. This is still a wide *open domain* for further research.

11. In our opinion, exactly the serious applications of information theory to fields other than the classical communication theory could make good use of this new, mixed theory of information. Indeed, in the classical communication theory the *contents* of messages are usually ignored and only their frequencies, probabilities, etc. considered. This makes the probabilistic information theory the right tool there. But it seems to us that in many other applications the contents are essential, and so it would be worth exploring whether the new theory could be more appropriately and efficiently applied there.

I did not aim at stating results in their most general form or at completeness in any sense (not even in the references). Also the opinions expressed here are subjective and some (several?) may turn out to be erroneous. But I hope that I have succeeded to give at least an idea of some non-communication applications of information theory or possibilities of such applications and also some food for thought concerning further work which seems worth doing.

Acknowledgements. This survey is based on a talk which the author presented at the Information Theory Workshop, Georgia Institute of Technology, April 15, 1982. An earlier version is appearing in the proceedings of that workshop in Information Processing & Management.

This research has been supported in part by the Natural Sciences and Engineering Research Council of Canada nr. A-2972.

REFERENCES

[1] ACZÉL, J., *Determination of all additive quasiarithmetic mean codeword lengths.* Z. Wahrsch. Verw. Gebiete *29* (1974), 351–360.

[2] ACZÉL, J., *A mixed theory of information—II: Additive inset entropies (of randomized systems of events) with measurable sum property.* Utilitas Math. *13* (1978), 49–54.

[3] ACZÉL, J., *A mixed theory of information. V. How to keep the (inset) expert honest.* J. Math. Anal. Appl. *75* (1980), 447–453.

[4] ACZÉL, J., *A mixed theory of information. VI. An example at last: A proper discrete analogue of the continuous Shannon measure of information (and its characterization).* Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No. 602–633 (1978–80), 65–72.

[5] ACZÉL, J., *A mixed theory of information. VII. Inset information functions of all degrees.* C.R. Math. Rep. Acad. Sci. Canada *2* (1980), 125–129.

[6] ACZÉL, J., *Functions partially constant on rings of sets.* C.R. Math. Rep. Acad. Sci. Canada *2* (1980), 159–164.

[7] ACZÉL, J. and DARÓCZY, Z., *Sur la caractérisation axiomatique des entropies d'ordre positif, comprise l'entropie de Shannon.* C.R. Acad. Sci. Paris *257* (1963), 1581–1584.

[8] ACZÉL, J. and DARÓCZY, Z., *Über verallgemeinerte quasilineare Mittelwerte, die mit Gewichtsfunktionen gebildet sind.* Publ. Math. Debrecen *10* (1963), 171–190.

[9] ACZÉL, J. and DARÓCZY, Z., *On measures on information and their characterizations.* Academic Press, New York–San Francisco–London, 1975.

[10] ACZÉL, J. and DARÓCZY, Z., *A mixed theory of information. I. Symmetric, recursive and measurable entropies of randomized systems of events.* RAIRO Inform. Théor. *12* (1978), 149–155.

[11] ACZÉL, J., FORTE, B., and NG, C. T., *Why the Shannon and Hartley entropies are "natural".* Adv. in Appl. Probab. *6* (1974), 131–146.

[12] ACZÉL, J. and KANNAPPAN, PL., *A mixed theory of information. III. Inset entropies of degree ß.* Inform. and Control *39* (1978), 315–322.

[13] ACZÉL, J. and NG, C. T., *Determination of all semisymmetric recursive information measures of multiplicative type on n positive discrete probability distributions.* Linear Algebra Appl. *52–53* (1983), 1–30.

[14] CAMPBELL, L. L., *A coding theorem and Rényi's entropy.* Inform. and Control *8* (1965), 423–429.

[15] CAMPBELL, L. L., *Definition of entropy by means of a coding problem.* Z. Wahrsch. Verw. Gebiete *6* (1966), 113–118.

[16] CSISZÁR, I., *Information measures: A critical survey.* In *Trans. Seventh Prague Conf. Information Theory, Statist., Dec. Functions, Random Processes and Eighth European Meeting of Statisticians.* Vol. B, Academia, Prague, 1978, pp. 73–86.

[17] DARÓCZY, Z., *Über Mittelwerte und Entropien vollständiger Wahrscheinlichkeitsverteilungen.* Acta Math. Acad. Sci. Hungar. *15* (1964), 203–210.

[18] DARÓCZY, Z., *Generalized information functions.* Inform. and Control *16* (1970), 36–51.

[19] DARÓCZY, Z., *On the measurable solutions of a functional equation.* Acta Math. Acad. Sci. Hungar. *22* (1971), 11–14.

[20] DARÓCZY, Z. and KÁTAI, I., *Additive zahlentheoretische Funktionen und das Mass der Information.* Ann. Univ. Sci. Budapest. Eötvös Sect. Math. *13* (1970), 83–88.

[21] DEVIJVER, P. A., *Entropie quadratique et reconnaissance des formes.* In *Computer Oriented Processes*, Noordhof, Leyden, 1976, pp. 257–277.

[22] DEVIJVER, P. A., *Entropies of degree β and lower bounds for the average error rate.* Inform. and Control *34* (1977), 222–226.

[23] DIDERRICH, G., *Local boundedness and the Shannon entropy.* Inform. and Control *29* (1975), 149–161.

[24] FISCHER, P., *On the inequality* $\sum p_i f(p_i) \geqq \sum p_i f(q_i)$. Metrika *18* (1972), 199–208.

[25] FISCHER, P., *On the inequaility* $\sum_{i=1}^{n} p_i f(p_i)/f(q_i) \leqq 1$. Canad. Math. Bull. *17* (1974), 193–199.

[26] FISCHER, P., *On the inequality* $\sum_{i=1}^{n} p_i f(p_i)/f(q_i) \geqq 1$. Pacific J. Math. *60* (1975), 65–74.

[27] FORTE, B., *Measures of information: The general axiomatic theory.* Rev. Française Informat. Recherche Opérationnelle *3* (1969), Sér. R-2, 63–89.

[28] GUIAŞU, S., *Information theory with applications.* McGraw-Hill International. New York–Auckland–Bogotá, 1977.

[29] HAVRDA, J. and CHARVÁT, F., *Quantification method of classification processes. Concept of structural a-entropy.* Kybernetika (Prague) *3* (1967), 30–35.

[30] KANNAPPAN, PL., *On generalizations of some measures in information theory.* Glasnik Mat. *9* (*29*) (1974), 81–93.

[31] KANNAPPAN, PL., *A mixed theory of information. IV. Inset inaccuracy and directed divergence.* Metrika *27* (1980), 91–98.

[32] KANNAPPAN, PL., and SANDER, W., *A mixed theory of information. VIII. Inset measures depending upon several distributions.* Aequationes Math. *25* (1982–83), 177–193.

[33] KAPUR, J. N., *Information of order α and type β.* Proc. Indian Acad. Sci. Sect. A *68* (1968), 65–75.

[34] KOLMOGOROV, A. N., *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin, 1933.

[35] MAKSA, GY., *Bounded symmetric information functions.* C.R. Math. Rep. Acad. Sci. Canada *2* (1980), 247–252.

[36] MEGINNIS, J. R., *A new class of symmetric utility rules for gambles, subjective marginal probability functions and a generalized Bayes rule.* Bus. Econom. Statist. Sec. Proc. Amer. Statist. Assoc. *1976*, 471–476.

[37] NG, C. T., *Representation for measures of information with the branching property.* Inform. and Control *25* (1974), 45–56.

[38] RÉNYI, A., *On measures of entropy and information.* In *Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1960*, Vol. 1, Univ. of Calif. Press, Berkeley, 1961, pp. 547–561.

[39] RÉNYI A., *On the foundations of information theory.* Rev. Inst. Internat. Statist. *33* (1965), 1–14.

[40] SHANNON, C. E. and WEAVER, W., *The mathematical theory of communication.* Univ. of Ill. Press, Urbana, 1949.

[41] SHARMA, B. D. and MITTAL, D. P., *New non-additive measures for discrete probability distributions.* J. Math. Sci. *10* (1975), 28–40.

[42] SHARMA, B. D. and TANEJA, I. J., *Entropy of type (α, β) and other generalized measures in information theory.* Metrika *22* (1975), 205–215.

[43] THEIL, H., *Economics and information theory.* North Holland, Amsterdam—Rand McNally, Chicago, 1967.

[44] VAN DER PYL, TH., *Axiomatique de l'information d'ordre α et de type β.* C.R. Acad. Sci. Paris Sér. A *28* (1976), 1031–1033.

[45] *The twentieth international symposium on functional equations, August 1–7, 1982 Oberwolfach, Germany* (compiled by B. Ebanks). Aequationes Math. *24* (1982), 261–297.

Centre for Information Theory
University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1