# INFORMATION MEASURES: A CRITICAL SURVEY

I. CSISZÁR

BUDAPEST*

## 1. INTRODUCTION

A key feature of Shannon's information theory is the discovery that the colloquial term information can often be given a mathematical meaning as a numerically measurable quantity, on the basis of a probabilistic model, in such a way that the solution of many important problems of information storage and transmission can be formulated in terms of this measure of the amount of information. This information measure has a very concrete operational interpretation: roughly, it equals the minimum number of binary digits needed, on the average, to encode the message in question. The coding theorems of information theory provide so overhelming evidence for the adequatness of Shannon's information measure that to look for essentially different measures of information might appear to make no sense at all. Moreover, it has been shown by several authors, starting with Shannon [35], that the measure of the amount of information is uniquely determined by some rather natural postulates. Still, all the evidence that Shannon's information measure is the only possible one, is valid only within the restricted scope of coding problems considered by Shannon. As Rényi pointed out in his fundamental paper [32] on generalized information measures, in other sorts of problems other quantities may serve, just as well or even better as measures of information. This should be indicated either by their operational significance (pragmatic approach) or by a set of natural postulates characterizing them (axiomatic approach) or, preferably, by both. In this mainly expository paper, some functionals of (complete) probability distributions meeting both criteria will be discussed, with a tendency of giving priority to the operational significance (which, however, need not be connected necessarily with coding).

Some new contributions included in the paper are the postulational characterization of $f$-divergences in Section 3 and the tentative application of $f$-informativity to estimate the reliability function of noisy channels in Section 4.

* This work was done while the author was visiting professor at Bowling Green State University, Bowling Green, Ohio, U.S.A.

For an exhaustive treatment of the axiomatic theory of information measures from the point of view of functional equations, see Aczél-Daróczy [1]. It should be noted that interesting axiomatic theories of information not based on the concept of probability also exist but, as far as this author knows, they still have to meet the "pragmatic" criterion.

There is an operationally justified information measure which will not be discussed in this paper since it arises in a rather special context. This is Fisher's information, familiar to statisticians for a long time. Let us remark, however, that it can be derived from the information measures considered in Section 3, see Kullback [27] and Vajda [38].

## 2. SHANNON'S AND RÉNYI'S INFORMATION MEASURES

Let $P = (p_1, \ldots, p_k)$ be a discrete PD (probability distribution). Let us designate by $X$ a random variable with distribution $P$ and by $Y$ another random variable; set $\boldsymbol{P}\{X = x_i,\ Y = y_j\} = r_{ij}$, $1 \leqq i \leqq k$, $1 \leqq j \leqq l$ and $P_i = (r_{i1}/p_i, \ldots, r_{il}/p_i)$, $1 \leqq i \leqq k$. If $X$ and $Y$ are independent, this will be designated by $X$ ind $Y$.

Shannon's measure of the average amount of information obtained when observing $X$, or of the uncertainty before this observation, is the entropy

$$H(P) = H(X) = -\sum_i p_i \log_2 p_i \,. \tag{2.1}$$

The conditional entropy defined as

$$H(Y \mid X) = \sum_i p_i H(P_i) = -\sum_{i,j} r_{ij} \log_2 \frac{r_{ij}}{p_i} \tag{2.2}$$

has a similar interpretation, and the mutual information

$$I(X; Y) = H(Y) - H(Y \mid X) = \sum_{i,j} r_{ij} \log_2 \frac{r_{ij}}{p_i q_j} \quad (q_j = \sum_i r_{ij}) \tag{2.3}$$

is a measure of the amount of information obtained from observing $X$ with respect to $Y$.

Some standard properties of Shannon's entropy with obvious intuitive meaning are

$$H(Y \mid X) \leqq H(Y), \quad \text{equality iff } X \text{ ind } Y, \tag{2.4}$$

$$H(X, Y) = H(X) + H(Y \mid X) \quad \text{(strong additivity)} \tag{2.5}$$

and their consequences

$$H(X, Y) \leqq H(X) + H(Y) \quad \text{(subadditivity)}, \tag{2.6}$$

$$H(X, Y) = H(X) + H(Y) \quad \text{if } X \text{ ind } Y \text{ (weak additivity)}. \tag{2.7}$$

Also, for the mutual information

$$I(X; Y) = I(Y; X) \geqq 0, \quad \text{equality iff } X \text{ ind } Y. \tag{2.8}$$

The simplest operational justification of the interpretation of Shannon's entropy as information measure is provided by the elementary coding theorems for memoryless sources, one for variable length codes (requiring exact decodability) and another for fixed length codes (requiring decodability with arbitrary small probability of error). In both cases, the entropy is the infimum of the achievable coding rates (average number of binary code digits per message symbol).

The true significance of Shannon's information measure becomes apparent, however, in the more complex problem of reliable information transmission over noisy channels. Here the mutual information plays the key role; we return to this point in Section 4.

In contrast to the "pragmatic" approach concentrating on the operational significance of the measure of information, the "axiomatic" approach starts from some a prirori desirable properties such as (2.4)–(2.7), and looks for the possible functions of PD's possessing them. The most standard postulates are that $H(p_1, \ldots, p_k)$ be a continuous, symmetric function of the $p_i$'s for every $k \geqq 2$, normed by $H(\frac{1}{2}, \frac{1}{2}) = 1$, and satisfying the following specialization of (2.5):

$$H(tp_1, (1-t)\, p_1, p_2, \ldots, p_k) = H(p_1, \ldots, p_k) + p_1\, H(t, 1-t) \quad (0 < t < 1); \tag{2.9}$$

of course, (2.5) itself is a direct consequence of (2.9) using symmetry.

By a theorem of Fadeev [15], these postulates imply that $H(p_1, \ldots, p_k)$ is of form (2.1). As a point of mathematical interest, the assumption of continuity can be essentially weakened; it suffices to postulate only Lebesgue measurability, as shown by Lee [28]. A very interesting recent development is Forte's result [16] (for a further improvement, see [2]) according to which Shannon's entropy remains the only possible information measure even if instead of strong additivity only subadditivity and weak additivity are postulated.

To substanciate the significance of this achievement of the axiomatic approach also for those whose interest is centered mainly on the pragmatic side, let us mention that in a very remarkable recent work of Ahlswede and Körner [3] on complex coding problems just the subadditivity and weak additivity of the entropy function were essentially used. If some other function with the same properties could have been found, their methods might have lead to strong rather than weak converses to the

proved coding theorems. Forte's theorem shows, however, that this way is not feasible.

If only weak additivity is required, there already exist functions of PD's other than Shannon's entropy with this property such as

$$H_\alpha(P) = \frac{1}{1-\alpha} \log_2 \sum_i p_i^\alpha \quad (0 < \alpha < 1 \text{ or } \alpha > 1). \tag{2.10}$$

The quantities (2.10) have been introduced by Schützenberger [34] and extensively studied by Rényi, see e.g. [32]. Rényi has called them entropies of order $\alpha$; they include Shannon's entropy in a limiting sense, namely $H_\alpha(P) \to H(P)$ as $\alpha \to 1$. For this reason, Shannon's entropy may be called entropy of order 1. Rényi's main motivation for considering these generalized information measures appears to be that he wanted to use them for proving limit theorems following an idea of Linnik [29]. Later he has demonstrated that $\alpha$-entropies naturally occur in the solution of certain search problems [33], and Campbell [6] has shown that the variable-length version of the elementary coding theorem carries over to $\alpha$-entropies, if in the definition of average code length one considers exponential averaging instead of the standard arithmetic one. $\alpha$-entropies do have some significance for fixed length codes, too: for the optimum binary encoding with rate $R > H(P)$ of messages of length $n$ from a memoryless source with distribution $P$, the error probability $p_n(e)$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log_2 \frac{1}{p_n(e)} = \max_{0<\alpha<1} \frac{1-\alpha}{\alpha} (R - H_\alpha(P)). \tag{2.11}$$

This is an equivalent formulation of a theorem of Jelinek [23]. Of course, since $H_\alpha(P)$ is in a functional relationship with the moment-generating function of the "information density" $I_X$ (defined by $I_X = -\log_2 p(X)$, $p(x) = P\{X = x\}$), every formula involving this moment-generating function can be rewritten in terms of $\alpha$-entropies.

Clearly, entropies of order $\alpha$ do have a reasonable operational significance even if not one comparable with that of Shannon's entropy. As regards the axiomatic approach, Rényi [32] did suggest a set of postulates characterizing his entropies but it involved the rather artificial device of considering incomplete distributions ($\sum p_i < 1$) as well. This shortcoming has been eliminated by Daróczy [13]; his main postulate in addition to weak additivity is that the entropy be of form $\psi^{-1}[\sum_i p_i \psi(-\log_2 p_i)]$ with some continuous, strictly monotonic function $\psi$.

For operational purposes, it seems more natural to consider instead of Rényi's entropy the simpler expression $\sum_i p_i^\alpha$ as an information measure, as e.g. Havrda and Charvát did in [21] (up to a constant factor). This quantity permits simpler postulational characterizations, too, see [14], [17], [21]. The characterization given by Forte and Ng [17] seems the most satisfactory; their main postulates are that $H(X, Y)$ be a function of $H(X)$ and $H(Y)$ if $X$ ind $Y$, and that $H(tp_1, (1-t)p_1, p_2, \ldots, p_k) - H(p_1, \ldots, p_k)$ be a function of $t$ and $p_1$, a weakening of Fadeev's postulate (2.9).

Quantities depending on two or more parameters and reducing in special cases to Rényi's or Shannon's entropy have also been proposed as generalized information measures. Since at present they do not have any operational significance, we omit mentioning them.

A further operationally justified entropy concept will be discussed in Section 4.

The concept of entropy (both Shannon's and Rényi's) has a remarkable shortcoming: it cannot be naturally extended to the non-discrete case. The obvious reason is that a random variable with a continuous distribution cannot be described by a finite number of binary digits thus it should have infinite entropy. In a sense, it is still meaningful to define the (Shannon) entropy of a random variable $X$ taking values in an arbitrary measurable space with respect to a given measure $\lambda$ on the same space by

$$H_\lambda(X) = -\int p \log_2 p \, d\lambda \tag{2.12}$$

where $p = p(x)$ is the density function of $X$ with respect to $\lambda$ (if the distribution of $X$ is not absolutely continuous with respect to $\lambda$, we set $H_\lambda(X) = -\infty$). The generalized Shannon entropy (2.12) measures the average information content of $X$ "apart from an infinitely large additive constant" in the following sense: For a partition $\mathscr{A}$ of the range of $X$, let $X^{\mathscr{A}}$ be a corresponding discrete approximation of $X$ (for each atom $A$ of $\mathscr{A}$, let $X^{\mathscr{A}}$ equal a fixed element of $A$ if $X \in A$); let $\delta(\mathscr{A})$ be the supremum of the diameters of the atoms of $\mathscr{A}$ (here the range of $X$ is assumed to be a metric space). Then, for partitions with atoms of equal interest, i.e., of equal $\lambda$-measure $\varepsilon = \varepsilon(\mathscr{A})$, we have

$$\lim_{\delta(\mathscr{A})\to 0} \left\{ H(X^{\mathscr{A}}) - \log \frac{1}{\varepsilon(\mathscr{A})} \right\} = H_\lambda(X)\,, \tag{2.13}$$

under a weak regularity condition. If $X$ is an $n$-dimensional random vector and $\lambda$ the $n$-dimensional Lebesgue measure, the regularity condition is fulfilled if $H([X]) < \infty$ where $[X]$ denotes the vector of the integer parts of the components of $X$ (corollary of theorem 1 of Csiszár [11]). A similar statement holds for the entropy of order $\alpha$, as well.

As indicated above, no direct coding justification of entropies of non-discrete random variables can be hoped for. The main importance of generalized Shannon entropy (2.12) lies in the fact that it can be used to calculate mutual information to the analogy of (2.3). Another pragmatic approach leading to the generalized Shannon or Rényi entropy $H_\lambda(X)$ or $H_{\lambda;\alpha}(X)$ is the following (Csiszár [11, theorem 3]): If the range of $X$ is quantized by a finite partition $\mathscr{A} = (A_1, \ldots, A_m)$, let the average quantization loss be measured by a mean value of the $\lambda(A_i)$'s of form

$$M_\beta(\mathscr{A}) = \{ \sum_{i=1}^{m} P\{X \in A_i\} \, \lambda^\beta(A_i) \}^{1/\beta} \quad (\beta > 0) \tag{2.14}$$

or

$$M_0(\mathscr{A}) = \exp_2 \{ \sum_{i=1}^{m} P\{X \in A_i\} \log_2 \lambda(A_i) \} ; \tag{2.15}$$

here the range of $X$ is assumed to have finite $\lambda$-measure.

Then, if $\lambda$ and the distribution of $X$ are non-atomic measures, the minimum-loss $m$-quantizations $\mathscr{A}_m$ satisfy

$$\lim_{m \to \infty} \{ \log_2 M_\beta(\mathscr{A}_m) + \log_2 m \} = H_{\lambda;\alpha}(X), \quad \alpha = \frac{1}{1+\beta} \tag{2.16}$$

provided that $\delta(\mathscr{A}_m) \to 0$.

Postulational characterizations of generalized Shannon entropy have also been suggested. Here we do not enter this question, only mention that the most elegant result of this type is due to Fritz [18].

## 3. INFORMATION-TYPE MEASURES OF DISTINGUISHABILITY OF PROBABILITY DISTRIBUTION

Let $P$ and $Q$ be PD's (measures with total mass 1) on an arbitrary measurable space and suppose that the distribution of a random variable $X$ is either $P$ or $Q$. The quantity

$$I(P \parallel Q) = \begin{cases} \int \log_2 \frac{dP}{dQ} \, dP = \int \frac{dP}{dQ} \log_2 \frac{dP}{dQ} \, dQ & \text{if } P \ll Q, \\ +\infty & \text{if } P \not\ll Q \end{cases} \tag{3.1}$$

introduced by Kullback and Leibler [26] will be referred to as $I$-divergence. It is nonnegative and vanishes only if $P = Q$.

A standard result attributed to Stein (see Chernoff [7]) asserts that testing the hypothesis $P$ against the alternative $Q$ from $n$ independent observations on $X$, if the best test of any fixed level $0 < \alpha < 1$ is used, the probability of second kind error satisfies

$$\lim_{n \to \infty} \frac{1}{n} \log_2 \frac{1}{\beta_n} = I(P \parallel Q). \tag{3.2}$$

Hence, $I(P \parallel Q)$ is an appropriate measure of statistical distinguishability of $P$ and $Q$. It should be emphasized that the $I$-divergence is not a metric (it is non-symmetric and the triangle inequality is not true, either); still, it is possible to formulate certain "geometric" propositions for PD's, the $I$-divergence playing the role of squared Euclidean distance, see Csiszár [12].

$I(P \parallel Q)$ can also be interpreted as the measure of the average information provided by one observation on $X$ for discriminating in favor of $P$ against the alternative $Q$, if $P$ is the true distribution. Though this information measure is conceptually different from Shannon's, there is a close relationship between them. Formally, $I(P \parallel Q)$ is minus the generalized entropy of $P$ with respect to $Q$, see (2.12). What is more essential, the mutual information of two random variables $X$ and $Y$ with distributions $P_X$, $P_Y$ and joint distribution $P_{XY}$ can be represented as

$$I(X; Y) = I(P_{XY} \parallel P_X \times P_Y)\,. \tag{3.3}$$

More exactly, in the discrete case (3.3) is clearly equivalent to (2.3) while in the general case (3.3) may serve as the definition of mutual information. The $I$-divergence plays an important role in coding theory not only through eq. (3.3) but on its own right, as well. Namely, it can be effectively used to describe the asymptotic behaviour of error probability both in channel and source coding; a well readable exposition of this method is Omura's paper [30].

Concerning postulational characterizations of the $I$-divergence, see e.g. [24].

Returning to the problem of testing whether $P$ or $Q$ is the true distribution of $X$, it is intuitively clear that replacing $X$ by a statistic $Y = T(X)$ one cannot increase the available information. Indeed, as shown already by Kullback-Leibler [26], we have

$$I(PT^{-1} \parallel QT^{-1}) \leqq I(P \parallel Q) \tag{3.4}$$

with equality iff $T$ is sufficient statistic for the pair $(P, Q)$ (or the left hand side of (3.4) is infinite).

A broad family of information-type measures of difference of PD's having the above intuitive property has been introduced by Csiszár [8], [9], and independently by Ali and Silvey [4] (and rediscovered by Zakai and Ziv [39]). If $f(t)$, $0 < t < \infty$ is any convex function, the $f$-divergence of $P$ and $Q$ is defined as

$$I_f(P \parallel Q) = \int q\, f\left(\frac{p}{q}\right) \mathrm{d}\lambda \tag{3.5}$$

where $p$ and $q$ are the densities of $P$ and $Q$ with respect to a dominating measure $\lambda$ (the choice of $\lambda$ is clearly irrelevant). In (3.5), undefined expressions are understood as $f(0) = \lim_{t \to 0} f(t)$, $0 f(0/0) = 0$, $0 f(p/0) = p \lim_{t \to 0} t f(1/t)$. The choice $f(t) = t \log_2 t$ gives the $I$-divergence, $f(t) = (t - 1)^2$ the $\chi^2$-divergence

$$\chi^2(P \parallel Q) = \int \frac{(p - q)^2}{q}\, \mathrm{d}\lambda \tag{3.6}$$

and $f(t) = |t - 1|$ the variation distance

$$|P - Q| = \int |p - q|\, \mathrm{d}\lambda\,. \tag{3.7}$$

More generally than (3.4), if $\Pi = \{P_x\}$ is any observation channel, i.e. a family of PD's such that $P_x(B)$ is a measurable function of $x$ for every measurable $B$, the output PD's $P\Pi$ and $Q\Pi$ defined by $(P\Pi)(B) = \int P_x(B)\,dP$, $(Q\Pi)(B) = \int P_x(B)\,dQ$ always satisfy

$$I_f(P\Pi \parallel Q\Pi) \leqq I_f(P \parallel Q)\,; \tag{3.8}$$

if $f$ is strictly convex, the equality holds iff $\Pi$ is a "sufficient channel" with respect to the pair $(P, Q)$, see [8], [9]. This property can be used for proving ergodicity of Markov processes, as done first by Rényi [32] and in a more general context by Kendall [25]; the limit theorem for convolution powers of PD's on a compact group has also been proved in this way, see Csiszár [8a]. An interesting further development of this method, for reversible Markov processes with general state space, is due to Fritz [19]; unlike the previously mentioned papers, he uses specifically the $I$-divergence.

The $f$-divergences can be used for estimation, as well. Suppose that $X$ is a discrete random variable and its distribution is known to belong to a convex set $\mathscr{E}$ of PD's $P = (p_1, \ldots, p_k)$. If the empirical distribution in a sample of $n$ independent observations on $X$ is $Q = (n_1/n, \ldots, n_k/n)$, one can choose that $P \in \mathscr{E}$ as an estimation of the distribution of $X$ which minimizes

$$I_f(P \parallel Q) = \sum_i \frac{n_i}{n} f\left(\frac{np_i}{n_i}\right). \tag{3.9}$$

Convexity ensures that the minimizing $P$ is unique (if $f$ is strictly convex).

For $f(t) = -\log_2 t$ we have

$$I_f(P \parallel Q) = \sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n} - \frac{1}{n} \log_2 \prod_i p_i^{n_i}\,, \tag{3.10}$$

hence minimization with respect to $P \in \mathscr{E}$ gives just the maximum likelihood estimator of $P$. Another familiar method, based on the $\chi^2$-statistic, is obtained by the choice $f(t) = (t-1)^2$, while $f(t) = t \log_2 t$ gives the minimum discrimination information method advocated by Kullback [27]. Observe that the maximum likelihood method is also equivalent with an $I$-divergence minimization, but with that of $I(Q \parallel P)$ rather than of $I(P \parallel Q)$ where $Q$ is the empirical distribution.

The $f$-divergence minimization method gives best asymptotically normal estimates of $P$, for any choice of $f$, as long as $f(t)$ is twice differentiable. This follows, e.g., from [36]. Thus, it is not possible to pick an optimal $f$ on the basis of asymptotic properties; unfortunately, little is known of the small sample properties. From a computational point of view, the $\chi^2$-method might seem the simplest, but as demonstrated e.g. by Ireland and Kullback [22], the $I$-divergence method often has a definite computational edge. The reason is that in most cases of interest, the constrained minimization problem can be solved numerically by a simple iterative

algorithm; a general formulation and convergence proof of this algorithm is given in Csiszár [12].

One theoretical advantage of the minimum $I$-divergence method when compared with the more standard maximum likelihood or minimum $\chi^2$-method is that it admits a successive adjustment of the estimator to additional contraints: in case of $\mathscr{E}_1 \supset \mathscr{E}_2$ where $\mathscr{E}_1$ is a linear manifold, if $P_1$ minimizes $I(P \parallel Q)$ for $P \in \mathscr{E}_1$ and $P_2$ minimizes $I(P \parallel P_1)$ for $P \in \mathscr{E}_2$, it follows that $P_2$ also minimizes $I(P \parallel Q)$ for $P \in \mathscr{E}_2$ (see Csiszár [12]).

Concerning other statistical applications of $f$-divergences see e.g. Perez [31].

Now we turn to an axiomatic approach and prove the following.

THEOREM 1. Let $F(P \parallel Q)$ be defined and finite valued for discrete PD's consisting of the same number of probabilities $p_i > 0$ and $q_i > 0$ $(i = 1, \ldots, k;\ k$ arbitrary). Suppose that

(i) $F(P \parallel Q)$ is invariant under simultaneous permutations of the $p_i'$s and $q_i'$s;

(ii)

$$F(p_1 + p_2, p_3, \ldots, p_k \parallel q_1 + q_2, q_3, \ldots, q_k) \leqq F(p_1, \ldots, p_k \parallel q_1, \ldots, q_k)$$

with equality if $p_1/q_1 = p_2/q_2$;

(iii)

$$F(tp_1', \ldots, tp_k', (1-t)\,p_1'', \ldots, (1-t)\,p_l'' \parallel tq_1', \ldots, tq_k', (1-t)\,q_1'', \ldots, (1-t)\,q_l'') =$$
$$= t\,F(p_1', \ldots, p_k' \parallel q_1', \ldots, q_k') + (1-t)\,F(p_1'', \ldots, p_l'' \parallel q_1'', \ldots, q_l'') \quad (0 < t < 1).$$

Then there exists a convex function $f$ such that $F(P \parallel Q) = I_f(P \parallel Q)$.

Here the postulates (i) and (ii) are particular cases of the property (3.4). Postulate (iii) requires that if two experiments can be performed with outcome governed by $P'$ and $P''$ under the null hypothesis and by $Q'$ and $Q''$ under the alternative hypothesis, then from the "mixed" experiment consisting in performing the first one with probability $t$ and the second with probability $1 - t$ we obtain information in favor of the null hypothesis against the alternative equal to the weighted average of the informations furnished by the two "pure" experiments.

Proof of Theorem 1. For given $P$ and $Q$, let $\mu$ be the atomic measure assigning mass $\mu(\{u\}) = \sum_{i:p_i/q_i=u} q_i$ to the possible values $p_i/q_i$ of $u$. Postulates (i) and (ii) imply that $F(P \parallel Q)$ depends only on $\mu$, i.e., $F(P \parallel Q) = F(\mu)$. Moreover, in view of postulate (iii),

$$F(t\mu' + (1-t)\,\mu'') = t\,F(\mu') + (1-t)\,F(\mu''). \tag{3.11}$$

For signed measures of form $\nu = \alpha\mu' - \beta\mu''$ $(\alpha \geqq 0,\ \beta \geqq 0)$, put $F(\nu) = \alpha\,F(\mu') - \beta\,F(\mu'')$. Using (3.11), one easily checks that $F(\nu)$ is uniquely defined and it is a linear functional on the linear space of the $\nu$'s of the said form. Extending this

functional to the linear space of all atomic signed measures on the positive half line, set $f(t) = F(\delta_t)$ where $\delta_t$ designates the unit mass concentrated at $t$. Then, using the linearity of the functional $F(\mu)$, we have

$$F(P \parallel Q) = F(\mu) = \sum_i q_i f\left(\frac{p_i}{q_i}\right) = I_f(P \parallel Q)\,. \tag{3.12}$$

It remains to check that $f(t)$ must be a convex function, but this is an immediate consequence of postulate (ii).

It should be noted that although postulate (iii) is a very natural one, there exist operationally justified information measures not fulfilling it. E.g., in hypothesis testing, if instead of the second kind error for fixed test level we are interested in the minimum of the sum of the errors of both kinds, then the role of $I(P \parallel Q)$ in (3.2) will be played by the Chernoff information number

$$D(P, Q) = \sup_{0<\alpha<1} \left\{-\log_2 \int p^\alpha q^{1-\alpha}\, d\lambda\right\}, \tag{3.13}$$

cf. Chernoff [7]. The latter information measure plays an important role also in problems with more than two hypotheses, see e.g. Vajda [37].

On the basis of Theorem 1, it is easy to get a new axiomatic characterization of the $I$-divergence; to this end, it suffices to add to (i)–(iii) a postulate of additivity.

## 4. GENERALIZATIONS OF THE MUTUAL INFORMATION

One of the most useful properties of Shannon's mutual information is the so called data processing theorem: if $W, X, Y, Z$ is a Markov chain then $I(W; Z) \leqq I(X; Y)$. It follows that if a message $W$ is to be transmitted over a channel within a prescribed distortion bound, the minimum mutual information needed to meet this bound (usually referred to as the rate-distortion function) must not exceed the maximum mutual information between the input and output of the channel called the channel capacity. The main coding theorems of information theory show that this bound is asymptotically tight, under reasonably general conditions, but it does not mean that shifting one's interest one cannot obtain better results when using some generalized information measure. As a generalization of Shannon's mutual information $I(X; Y)$ the first candidate is, to the analogy of (3.3) the $f$-divergence of $P_{XY}$ and $P_X \times P_Y$ (formula (2.3) seems less suitable for generalization, though Arimoto [5] did arrive at remarkable results in this way, using a new type of generalized entropy to be mentioned later). In view of (3.8), the data processing theorem is easily seen to remain true and certain consequences of statistical and/or information-theoretic significance

may be drawn. This approach was hinted by Csiszár [9] and more fully developped by Perez [31] and Zakai-Ziv [39].

Another approach which this author considers even more promising starts from the easily checked identity

$$I(P_{XY} \parallel P_X \times Q) = I(P_{XY} \parallel P_X \times P_Y) + I(P_Y \parallel Q) \tag{4.1}$$

where $Q$ is any PD on the range space of $Y$. Thus we have

$$I(X; Y) = I(P_{XY} \parallel P_X \times P_Y) = \min_Q I(P_{XY} \parallel P_X \times Q)\,. \tag{4.2}$$

This suggests to define, for any convex $f$,

$$I_f(X; Y) = \inf_Q I_f(P_{XY} \parallel P_X \times Q)\,. \tag{4.3}$$

In the case of discrete $X$, the properties of $I_f(X; Y)$ have been studied in detail in [10]. The data processing theorem is valid for $I_f(X; Y)$, and the latter quantity can be effectively used to characterize sufficiency of experiments with any finite number of permitted hypothetical distributions. An interesting feature of the quantity (4.3) is that in the case $f(t) = -t^\alpha$, $0 < \alpha < 1$, it becomes

$$I_\alpha(X; Y) = -\left\{\int \left(\sum_i w_i p_i^\alpha(y)\right)^{1/\alpha} \lambda(\mathrm{d}y)\right\}^\alpha \tag{4.4}$$

where $W = (w_1, \ldots, w_k)$ is the distribution of $X$ and $p_i(y)$ is the density (with respect to a dominating measure $\lambda$) of the conditional distribution of $Y$ given $X = x_i$.

Setting $\alpha = 1/(1 + \varrho)$, the function in the brackets in (4.4) equals $\exp E_0(\varrho, W)$, where $E_0(\varrho, W)$ is Gallager's function occuring in the reliability function of the noisy channel with transition probability densities $p_i(y)$, (see [20], p. 322). This indicates that the $\alpha$-informativity measures (4.4) might be used to prove e.g. the sphere-packing bound for the error probability in noisy channels. The method of proof would be to compare the minimum $I_\alpha$ needed to represent the message with error probability decreasing with a prescribed exponential rate and the maximum $I_\alpha$ permissible by the given channel. Unfortunately, the straightforward calculation leads to a weaker lower bound for the error probability than the sphere-packing bound. The reason is that the joint distribution of the original and decoded message minimizing $I_\alpha$ at given error probability corresponds to uniform conditional distribution of the decoded message in case of error. This cannot be realized over a memoryless channel, because, intuitively, at the decoder only a few codewords come really into account. This heuristic reasoning suggests a simple way of proving the sphere-packing bound which, however, still has to be made rigorous.

Let us remark that (4.3) can also be used to define a new concept of entropy, namely by

$$H_f(X) = I_f(X; X)\,. \tag{4.5}$$

This is related to the generalized entropy of Arimoto [5] who has defined

$$H_f(X) = \inf_Q \sum_i p_i f(q_i) \tag{4.6}$$

where $p_i = P\{X = x_i\}$. In fact, substituting $f$ in (4.5) by $\tilde{f}$ defined by $\tilde{f}(t) = t f(1/t)$, which is also convex if $f$ is, we obtain (4.6) provided that $\tilde{f}(0) = 0$.

Using his entropy (4.6), Arimoto has defined generalized mutual information according to (2.3) and obtained interesting applications of this concept.

In this section, only the pragmatic approach was used. The reason is the apparent lack of "true" postulational characterizations of mutual information, i.e. of characterizations not starting from eq. (2.3). To the knowledge of this author, an axiomatic approach to generalizations of Shannon's mutual information has not been attempted yet.

## REFERENCES


[1] J. Aczél, Z. Daróczy: On measures of information and their characterizations. Academic Press, New York 1974.

[2] J. Aczél, B. Forte, C. T.Ng: Why the Shannon and Hartley entropies are "natural". Advances in Applied Probability, *6* (1974), 131—146.

[3] R. Ahlswede, J. Körner: A source coding problem with side information and a converse for degraded broadcast channels. IEEE Transactions on Information Theory, *IT-27* (1975), 629—637.

[4] S. M. Ali, S. D. Silvey: A general class of coefficients of divergence of one distribution from another. J. Roy. Statist. Soc. B *28* (1966), 131—142.

[5] S. Arimoto: Information-theoretical consideration on estimation problems. Information and Control *19* (1971), 181—194.

[6] L. L. Campbell: A coding theorem and Rényi's entropy. Information and Control *8* (1965), 423—429.

[7] H. Chernoff: A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. Ann. Math. Statist. *23* (1952), 493—507.

[8] I. Csiszár: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Magyar Tud. Akad. Mat. Kutató Int. Közl. *8* (1963), 85—108.

[8a] I. Csiszár: A note on limiting distributions on topological groups. Magyar Tud. Akad. Mat. Kutató Int. közl. *9* (1964), 595—599.

[9] I. Csiszár: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. *2* (1967), 299—318.

[10] I. Csiszár: A class of measures of informativity of observation channels. Periodica Math. Hungar. *2* (1972), 191—213.

[11] I. Csiszár: Generalized entropy and quantization problems. In: Trans. of the Sixth Prague Conference, Prague 1971. Academia, Prague 1973, 159—174.

[12] I. CSISZÁR: $I$-divergence geometry of probability distributions and minimization problems. Annals of Probability, *3* (1975), 146—158.

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

[14] Z. DARÓCZY: Generalized information functions. Information and Control *16* (1970), 36—51.

[15] D. K. FADEEV: On the concept of entropy of a finite probabilistic scheme (in Russian). Uspechi Mat. Nauk. *11* (1956), 227—231.

[16] B. FORTE: Why Shannon's entropy. In: Covegno Inform. Teor. Ist. Naz. Alta Mat., Roma 1973. Symposia Math. Vol. 11, Academic Press, New York 1974.

[17] B. FORTE, C. T. NG: On a characterization of the entropies of type $\beta$. Utilitas Math. *4* (1973), 193—205.

[18] J. FRITZ: On the characteristic properties of generalized entropy. Problems of Control and Information Theory *1* (1972), 177—191.

[19] J. FRITZ: An information-theoretical proof of limit theorems for reversible Markov processes. In: Trans. of the Sixth Prague Conference, Prague 1971. Academia, Prague, 1973, 183—197.

[20] R. G. GALLAGER: Information theory and reliable communication. Wiley, New York 1968.

[21] J. HAVRDA, F. CHARVÁT: Quantification method of classification processes: Concept of structural $a$-entropy. Kybernetika *3* (1967), 30—34.

[22] C. T. IRELAND, S. KULLBACK: Contingency tables with given marginals. Biometrika *55* (1968), 179—188.

[23] F. JELINEK: Probabilistic information theory. Mc Graw Hill, New York 1968.

[24] PL. KANNAPAN, P. N. RATHIE: On various characterizations of directed-divergence. In: Trans. of the Sixth Prague Conference, Prague 1971. Academia Prague 1973, 331—339.

[25] D. G. KENDALL: Information theory and the limit theorem for Markov chains and processes with a countable infinity of states. Ann. Inst. Statist. Math. *15* (1963), 137—143.

[26] S. KULLBACK, R. A. LEIBLER: On information and sufficiency. Ann. Math. Statist. *22* (1951), 79—86.

[27] S. KULLBACK: Information theory and statistics. Wiley, New York 1959.

[28] P. M. LEE: On the axioms of information theory. Ann. Math. Statist. *35* (1964), 415—418.

[29] YU. V. LINNIK: Information-theoretic proof of the central limit theorem under Lindeberg's conditions (in Russian). Téor. Veroyatnost. i Primenen. *4* (1959), 311—321.

[30] J. K. OMURA: A lower bounding method for channel and source coding probabilities. Information and Control *27* (1975), 148—177.

[31] A. PEREZ: Risk estimates in terms of generalized $f$-entropies In: Proc. Colloquium on Information Theory, Debrecen 1967. J. Bolyai Mathematical Society, Budapest 1968, 299—315.

[32] A. RÉNYI: On measures of entropy and information. In: Proc. Fourth Berkeley Symposium 1960, Vol. 1. Univ. of California Press, Berkeley—Los Angeles 1961, 547—561.

[33] A. RÉNYI: On the foundations of information theory. Rev. Inst. Internat. Stat. *33* (1965), 1—14.

[34] M. P. SCHÜTZENBERGER: Contribution aux applications statistic de la théorie de l'information. Publ. Inst. Statist. Univ. Paris *3* (1954), 3—117.

[35] C. E. SHANNON: A mathematical theory of communication. Bell System Technical Journal *27* (1948), 379—423 and 623—656.

[36] W. T. TAYLOR: Distance functions and regular BAN estimates. Ann. Math. Statist. *24* (1953), 85—92.

[37] I. VAJDA: On the convergence of information contained in a sequence of observations. In: Proc. Colloquium on Information Theory. Debrecen 1967. J. Bolyai Mathematical Society, Budapest 1968.

[38] I. VAJDA: $\chi^{\alpha}$-divergence and generalized Fischer's information In: Trans. of the Sixth Prague Conference, Prague 1971. Academia, Prague 1973, 873—886.

[39] J. ZIV, M. ZAKAI: On functionals satisfying a data-processing theorem. IEEE Transactions *IT-19* (1973), 275—282.

HUNGARIAN ACADEMY OF SCIENCES

MATHEMATICAL INSTITUTE