Entropy **2008**, *10*, 261-273; DOI: 10.3390/e10030261

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.org/entropy

Article

Axiomatic Characterizations of Information Measures

Imre Csiszár

Rényi Institute of Mathematics, Hungarian Academy of Sciences, P.O.Box 127, H1364 Budapest, Hungary. E-mail: csiszar@renyi.hu

Received: 1 September 2008 / Accepted: 12 September 2008 / Published: 19 September 2008

Abstract: Axiomatic characterizations of Shannon entropy, Kullback I-divergence, and some generalized information measures are surveyed. Three directions are treated: (A) Characterization of functions of probability distributions suitable as information measures. (B) Characterization of set functions on the subsets of $\{1, \ldots, N\}$ representable by joint entropies of components of an N-dimensional random vector. (C) Axiomatic characterization of MaxEnt and related inference rules. The paper concludes with a brief discussion of the relevance of the axiomatic approach for information theory.

Keywords: Shannon entropy; Kullback I-divergence; Rényi information measures; f-divergence; f-entropy; functional equation; proper score; maximum entropy; transitive inference rule; Bregman distance

1. Introduction

Axiomatic characterizations of Shannon entropy

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i$$

and Kullback I-divergence (relative entropy)

$$D(P||Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i},$$

and of some generalized information measures will be surveyed, for discrete probability distributions $P = (p_1, \ldots, p_n)$, $Q = (q_1, \ldots, q_n)$, $n = 2, 3, \ldots$.

No attempt at completeness is made, the references cover only part of the historically important contributions, but are believed to be representative for the development of the main ideas in the field. It is also illustrated how a research direction originating in information theory developed into a branch of the theory of functional equations; the latter, however, is not entered in depth, for its major achievements appear to be solutions of mathematical problems beyond information-theoretic relevance.

1.1. Historical comments

"Shannon entropy" first appeared in statistical physics, in works of Boltzmann and Gibbs, in the 19th century. Quantum entropy, of a density matrix with eigenvalues $p_1, \ldots, p_n$, is defined by the same expression, Neumann [45]. I-divergence was defined as information measure by Kullback-Leibler [40] and may have been used much earlier in physics. The non-negativity of I-divergence is sometimes called Gibbs' inequality, but this author could not verify that it does appear in Gibbs' works. Wald [58] used I-divergence as a tool (without a name) in sequential analysis.

It was Shannon's information theory [52] that established the significance of entropy as a key information measure, soon complemented by I-divergence, and stimulated their profound applications in other fields such as large deviations [50], ergodic theory [38], and statistics [39].

Axiomatic characterizations of entropy also go back to Shannon [52]. In his view, this is "in no way necessary for the theory" but "lends a certain plausibility" to the definition of entropy and related information measures. "The real justification resides" in operational relevance of these measures.

1.2. Directions of axiomatic characterizations

(A) Characterize entropy as a function of the distribution $P = (p_1, \ldots, p_n)$, $n = 2, 3, \ldots$: Show that it is the unique function that satisfies certain postulates, preferably intuitively desirable ones. Similarly for I-divergence. This direction has an extensive literature. Main references: Aczél-Daróczy [1], Ebanks-Sahoo-Sander [26].

(B) Characterize entropy as a set function: Determine the class of set functions $\varphi(A)$, $A \subset \{1, \ldots, N\}$, which can be represented as $\varphi(A) = H(\{X_i\}_{i \in A})$, for suitable random variables $X_1, \ldots, X_N$, or as a limit of a sequence of such "entropic" set functions. This direction has been initiated by Pippenger [47], the main reference is Yeung [59].

(C) Characterize axiomatically the MaxEnt inference principle. To infer a distribution $P = (p_1, \ldots, p_n)$ from incomplete information specifying only linear constraints $\sum_{i=1}^{n} p_i a_{ij} = b_j$, $j = 1, \ldots, k$, this principle (Jaynes [33], Kullback [39]) calls for maximizing $H(P)$ or, if a "prior guess" Q is available, minimizing $D(P||Q)$ subject to the given constraints. References: Shore-Johnson [53], Paris-Vencovská [46], Csiszár [18].

(D) Not entered: Information without probability [32], [35], and the "mixed theory of information" [2].

2. Direction (A)

Properties of entropy that have been used as postulates:
– Positivity: $H(P) \geq 0$
– Expansibility: "Expansion" of P by a new component equal to 0 does not change $H(P)$
– Symmetry: $H(P)$ is invariant under permutations of $p_1, \ldots, p_n$
– Continuity: $H(P)$ is a continuous function of P (for fixed n)
– Additivity: $H(P \times Q) = H(P) + H(Q)$
– Subadditivity: $H(X, Y) \leq H(X) + H(Y)$
– Strong additivity: $H(X, Y) = H(X) + H(Y|X)$
– Recursivity: $H(p_1, \ldots, p_n) = H(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2)H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$
– Sum property: $H(P) = \sum_{i=1}^{n} g(p_i)$, for some function g.

Above, $H(X)$, $H(Y)$, $H(X, Y)$ are the entropies of the distributions of random variables X, Y (with values in $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$) and of their joint distribution. $H(Y|X)$ denotes the average of the entropies of the conditional distributions of Y on the conditions $X = i$, $1 \leq i \leq n$, weighted by the probabilities of the events $X = i$.

2.1. Shannon entropy and I-divergence

Shannon [52] showed that continuity, strong additivity, and the property that $H(1/n, \ldots, 1/n)$ increases with n, determine entropy up to a constant factor. The key of the proof was to show that the assumptions imply $H(1/n, \ldots, 1/n) = c \log n$.

Faddeev [27] showed that recursivity plus 3-symmetry (symmetry for $n = 3$) plus continuity for $n = 2$ determine $H(P)$ up to constant factor.

Further contributions along these lines include
Tverberg [56] and Lee [41]: relaxed continuity to Lebesgue integrability resp. measurability
Diderrich [25]: Recursivity plus 3-symmetry plus boundedness suffice
Daróczy-Maksa [24]: Positivity instead of boundedness does not suffice.

These works used as a key tool the functional equation for $f(x) = H(x, 1 - x)$

$$f(x) + (1 - x)f\left(\frac{y}{1 - x}\right) = f(y) + (1 - y)f\left(\frac{x}{1 - y}\right)$$

where $x, y \in [0, 1)$, $x + y \leq 1$. Aczél-Daróczy [1] showed that all solutions of this equation with $f(0) = f(1) = 0$ are given by

$$f(x) = xh(x) + (1 - x)h(1 - x) \quad 0 < x < 1,$$

where h is any function satisfying

$$h(uv) = h(u) + h(v) \quad u, v > 0.$$

Chaundy-McLeod [13] showed, by solving another functional equation, that the sum property with continuous g, plus additivity, determine Shannon's entropy up to a constant factor.

Daróczy [23] proved the same under the weaker conditions that g is measurable, $g(0) = 0$, and H is (3,2)-additive (additive for $P = (p_1, p_2, p_3)$, $Q = (q_1, q_2)$). However, (2,2)-additivity does not suffice.

The intuitively most appealing axiomatic result is due to Aczél-Forte-Ng [3], extending Forte's previous work [29]: Symmetry, expansibility, additivity, and subadditivity uniquely characterize linear combinations with non-negative coefficients of $H(P)$ and $H_0(P) = \log |\{i : p_i > 0\}|$. The same postulates plus continuity for $n = 2$ determine Shannon entropy up to a constant factor.

I-divergence has similar characterizations as entropy, both via recursivity, and the sum property plus additivity. For I-divergence, recursivity means

$$D(p_1, \dots, p_n||q_1, \dots, q_n) = D(p_1 + p_2, \dots, p_n||q_1 + q_2, \dots, q_n)+ \\ (p_1 + p_2)D\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}||\frac{q_1}{q_1 + q_2}, \frac{q_2}{q_1 + q_2}\right).$$

The sum property means that, for some function G of two variables,

$$D(p_1, \dots, p_n||q_1, \dots, q_n) = \sum_{i=1}^{n} G(p_i, q_i).$$

The first results in this direction employed the device of admitting also "incomplete distributions" for Q (with sum of probabilities less than 1). Not needing this, Kannappan-Ng [36-37] proved: $D(P||Q)$, as a function of arbitrary probability distributions P and strictly positive probability distributions Q, is determined up to a constant factor by recursivity, 3-symmetry, measurability in p for fixed q and in q for fixed p of $D(p, 1 - p||q, 1 - q)$, plus $D(1/2, 1/2||1/2, 1/2) = 0$.

For the proof the following analogue, with four unknown functions, of the functional equation in the characterization of entropy had to be solved:

$$f_1(x) + (1 - x)f_2\left(\frac{y}{1 - x}\right) = f_3(y) + (1 - y)f_4\left(\frac{x}{1 - y}\right),$$

where $x, y \in [0, 1)$, $x + y \leq 1$.

Both kinds of characterizations have been extended to "information measures" depending on more than two distributions. Here, only the following corollary of more general (deep) results in the book [26] is mentioned, as an illustration. If a function of m strictly positive probability distributions $P_j = (p_{j1}, \dots, p_{jn})$, $j = 1, \dots, m$, is of form $\sum_{i=1}^{n} G(p_{1i}, \dots, p_{mi})$ with measurable G, and satisfies additivity, then this function equals a linear combination of entropies $H(P_j)$ and divergences $D(P_i||P_j)$.

2.2. *Rényi entropies and divergences*

Shannon entropy and I-divergence are means $\sum p_k I_k$ of individual informations $I_k = -\log p_k$ or $I_k = \log \frac{p_k}{q_k}$. Rényi [48] introduced alternative information measures, which are generalized means $\psi^{-1}(\sum p_k \psi(I_k))$, where ψ is a continuous, strictly monotone function, and which satisfy additivity. Entropy and divergence of order $\alpha \neq 1$ correspond to $\psi(x)$ equal to $e^{(1-\alpha)x}$ respectively $e^{(\alpha-1)x}$:

$$H_\alpha(P) = \frac{1}{1 - \alpha} \log \sum p_k^\alpha, \quad D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \sum p_k^\alpha q_k^{1-\alpha};$$

here, the sums are for $k \in \{1, \ldots, n\}$ with $p_k > 0$. Limiting as $\alpha \to 1$ gives $H_1 = H$, $D_1 = D$.

These quantities were previously considered by Schützenberger [51]. Rényi [48] showed for the case of divergence, and conjectured also for entropy, that only these generalized means give rise to additive information measures, provided "incomplete distributions" were also considered. The latter conjecture was proved by Daróczy [20]. Then Daróczy [21] showed without recourse to incomplete distributions that the entropies of order $\alpha > 0$ exhaust all additive entropies equal to generalized means such that the entropy of $(p, 1-p)$ approaches 0 as $p \to 0$ (the last condition excludes $\alpha \leq 0$.)

Rényi entropies are additive, but not subadditive unless $\alpha = 1$ or 0. If $P = (1/n, \ldots, 1/n)$ then $H_\alpha(P) = \log n$, otherwise $H_\alpha(P)$ is a strictly decreasing function of α. Moreover, $H_\infty(P) := \lim_{\alpha\to\infty} H_\alpha(P) = -\log \max_k p_k$.

Rényi entropies have operational relevance in the theory of random search [49], for variable length source coding (average codelength in exponential sense [12]), block coding for sources and channels (generalized cutoff rates [19]), and in cryptography (privacy amplification [9]).

Remark 1. For information transmission over noisy channels, a key information measure is mutual information, which can be expressed via entropy and I-divergence in several equivalent ways. The α-analogues of these expressions are no longer equivalent, the one of demonstrated operational meaning is

$$I_\alpha(P, W) = \min_Q \sum_{k=1}^{n} p_k D_\alpha(W_k || Q),$$

see Csiszár [19]. Here W is the channel matrix with rows $W_k = (w_{k1}, \ldots, w_{km})$, $P = (p_1, \ldots, p_n)$ is the input distribution, and the minimization is over distributions $Q = (q_1, \ldots, q_m)$. This definition of mutual information of order α, and different earlier ones (Sibson [54], Arimoto [7]) give the same maximum over input distributions P ("capacity of order α" of the channel W).

2.3. *Other entropies and divergences*

The f-divergence of P from Q is

$$D_f(P\|Q) = \sum_{k=1}^{n} q_k f\left(\frac{p_k}{q_k}\right),$$

where f is a convex function on $(0, \infty)$ with $f(1) = 0$. It was introduced by Csiszár [14-15], and independently by Ali-Silvey [5]. An unsophisticated axiomatic characterization of f-divergences appears in Csiszár [17].

In addition to I-divergence, this class contains reversed I-divergence, Hellinger distance, χ^2-divergence, variation distance, etc. It shares some key properties of I-divergence, in particular monotonicity: for any partition $\mathcal{A} = (A_1, \ldots, A_m)$ of $\{1, \ldots, n\}$, with notation $P^{\mathcal{A}} = (p_1^{\mathcal{A}}, \ldots, p_m^{\mathcal{A}})$, $p_i^{\mathcal{A}} = \sum_{k\in A_i} p_k$, it holds that

$$D_f(P^{\mathcal{A}}\|Q^{\mathcal{A}}) \leq D_f(P\|Q).$$

Remark 2. The f-divergence of probability distributions does not change if $f(t)$ is replaced by $f(t) + a(t-1)$, any $a \in \mathbb{R}$, hence $f \geq 0$ may be assumed without loss of generality. If $f \geq 0$, the obvious extension of the definition to arbitrary $P, Q \in \mathbb{R}_+^n$ retains the intuitive meaning of divergence. Accordingly, the I-divergence of arbitrary $P, Q \in \mathbb{R}^n_+$ is defined as the f-divergence with $f(t) = t \log t - t + 1$,

$$D(P\|Q) = \sum_{i=1}^{n} (p_i \log \frac{p_i}{q_i} - p_i + q_i).$$

A generalization of mutual information via f-divergences, and as a special case a concept of *f-entropy*, appear in Csiszár [16]. Different concepts of f-entropies were defined by Arimoto [6], viz. $H^f(P) = \sum_{k=1}^n f(p_k)$, f concave, and $H_f(P) = \inf_Q \sum_{k=1}^n p_k f(q_k)$. Both were used to bound probability of error. Ben-Bassat [8] determined the best bounds possible in terms of $H^f(P)$. The f-entropy of [16] coincides with $H_{\tilde{f}}(P)$ in the sense of [6], where $\tilde{f}(x) = x f(1/x)$.

Very general information measures have been considered in the context of statistical decision theory, see Grünwald and Dawid [30] and references there. A function $l(Q, k)$ of probability distributions $Q = \{q_1, \ldots, q_n\}$ and $k \in \{1, \ldots, n\}$, measuring the loss when Q has been inferred and outcome k is observed, is called a *proper score* if the average loss $\sum_{k=1}^n p_k l(Q, k)$ is minimized for Q equal to the true distribution P, whatever this P is. Then $\sum_{k=1}^n p_k l(P, k)$ is called the entropy of P corresponding to the proper score l. In this context, Shannon entropy is distinguished as that corresponding to the only proper score of form $l(Q, k) = f(q_k)$, the logarithmic score. Indeed, if for some $n > 2$

$$\sum_{k=1}^{n} p_k f(p_k) \le \sum_{k=1}^{n} p_k f(q_k)$$

for all strictly positive distributions P and Q on $\{1, \ldots, n\}$, then $f(x) = c \log x + b$, with $c \le 0$. This result has a long history, the book [1] attributes its first fully general and published proof to Fischer [28].

In the decision theory framework, Arimoto's entropies $H^f(P)$ correspond to "separable Bregman scores" [30].

2.4. Entropies and divergences of degree α

This subclass of f-entropies/divergences is defined, for $\alpha \neq 0, 1$, by

$$H^\alpha(P) = c_\alpha \left(\sum_{k=1}^{n} p_k^\alpha - 1 \right), \quad D^\alpha(P\|Q) = -c_\alpha \left(\sum_{k=1}^{n} p_k^\alpha q_k^{1-\alpha} - 1 \right).$$

Here c_α is some constant, positive if $0 < \alpha < 1$ and negative otherwise. Its typical choices are such that $(1 - \alpha)c_\alpha \to 1$ as $\alpha \to 1$, then limiting as $\alpha \to 1$ gives $H^1 = H$, $D^1 = D$.

Entropy of degree α was introduced by Havrda-Charvát [31]. The special case of $\alpha = 2$ ("quadratic entropy") may have appeared earlier, Vajda [57] used it to bound probability of error for testing multiple hypotheses. Divergences of degree 2 and $1/2$ have long been used in statistics, the former since the early 20th century (χ^2 test), the latter goes back at least to Bhattacharyya [10].

In statistical physics, $H^\alpha(P)$ is known as Tsallis entropy, referring to [55]. Previously, Lindhard-Nielsen [42] proposed generalized entropies for statistical physics, effectively the same as entropies of degree α and order α, also unaware of their prior use in information theory.

Entropies/divergences of order α and those of degree α are in a one-to-one functional relationship. In principle, it would suffice to use only one of them, but in different situations one or the other is more convenient. For example, in source coding for identification it is entropy of degree 2 that naturally enters [4].

Entropies of degree $\alpha \geq 1$ are subadditive, but entropies of any degree $\alpha \neq 1$ are neither additive nor recursive. Rather,

$$H^\alpha(P \times Q) = H^\alpha(P) + H^\alpha(Q) + c_\alpha^{-1} H^\alpha(P) H^\alpha(Q),$$

$$H^\alpha(p_1, \dots, p_n) = H^\alpha(p_1 + p_2, p_3, \dots, p_n) + (p_1 + p_2)^\alpha H^\alpha \left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \right).$$

With these "α-additivity" and "α-recursivity", the analogues of characterization theorems for Shannon entropy hold, the first one due to Havrda-Charvát [31]. Remarkably, characterization via α-recursivity requires no regularity conditions [22]. Similar results hold for divergence of degree α, and for "information measures" of degree α involving more than two distributions. See the book [26] for details, some very complex. For divergence, α-recursivity means

$$D^\alpha(p_1, \dots, p_n \| q_1, \dots, q_n) = D^\alpha(p_1 + p_2, p_3, \dots, p_n \| q_1 + q_2, q_3 \dots, q_n) +$$
$$(p_1 + p_2)^\alpha (q_1 + q_2)^{1-\alpha} D^\alpha \left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \| \frac{q_1}{q_1 + q_2}, \frac{q_2}{q_1 + q_2} \right).$$

3. Direction (B)

This very important direction can not be covered here in detail. We mention only the following key results: For $N \geq 4$, the closure of the class of "entropic" set functions is a proper subclass of polymatroids, Zhang-Yeung [60]. It is a convex cone, Yeung [59], but not a polyhedral cone, Matúš [44], i.e., no finite set of linear entropy inequalities can provide the requested characterization.

4. Direction (C)

Here, some of the axiomatic results of Csiszár [18] are surveyed. Attention is not restricted to probability distributions, the object to be inferred could be
(i) a probability distribution $P = (p_1, \dots, p_n)$, or
(ii) any $P = (p_1, \dots, p_n) \in \mathbb{R}_+^n$, or
(iii) any $P \in \mathbb{R}^n$.
For technical reasons, in (i) and (ii) strict positivity of each p_k is required. This conforms with the intuitive desirability of excluding inferences that certain events have probability 0. Below, n is fixed, $n \geq 5$ in case (i), $n \geq 3$ in cases (ii), (iii).

The only information about P is that it belongs to a feasible set F which could be any nonempty set determined by constraints

$$\sum_{i=1}^{n} p_i a_{ij} = b_j, \quad j = 1, \dots, m,$$

that is, consisting of all P as in (i), (ii) or (iii) that satisfy the constraints. Assume that a prior guess (default model) Q is available which could be arbitrary (as P in (i), (ii) or (iii)).

An *inference rule* is any mapping Π that assigns to each feasible set F and prior guess Q an inference $\Pi(F, Q) = P^* \in F$. Axioms will be stated as desiderata for a "good" inference rule. The results substantiate that in cases (i) and (ii) the "best" inference rule is to let $\Pi(F, Q)$ be the I-projection of Q to F (MaxEnt), and that in case (iii) the regular Euclidean projection (least squares) is "best". Reasonable alternative rules will also be identified.

In the second axiom, we use the term "set of I-local constraints" where I is a subset of $\{1, \dots, n\}$. This means constraints of form $\sum_{i \in I} p_i a_{ij} = b_j$; in case (i) it is also supposed that one of them is $\sum_{i \in I} p_i = t$, for some $0 < t < 1$.

The axioms are as follows:

Regularity: (a) $Q \in F$ implies $\Pi(F, Q) = Q$, (b) $F_1 \subset F$ and $\Pi(F, Q) \in F_1$ imply $\Pi(F_1, Q) = \Pi(F, Q)$, (c) for each $P \neq Q$, among the feasible sets determined by a single constraint there exists a unique F such that $\Pi(F, Q) = P$, (d) $\Pi(F, Q)$ depends continuously on F.

Locality: If F_1 is defined by a set of I-local constraints, F_2 by a set of I^c-local ones, then the components p_i^*, $i \in I$ of $P^* = \Pi(F_1 \cap F_2, Q)$ are determined by F_1 and $\{q_i : i \in I\}$.

Transitivity: If $F_1 \subset F$, $\Pi(F, Q) = P^*$, then $\Pi(F_1, Q) = \Pi(F_1, P^*)$.

Semisymmetry: If $F = \{P : p_i + p_j = t\}$ for some $i \neq j$ and constant t, and Q satisfies $q_i = q_j$, then $P^* = \Pi(F, Q)$ satisfies $p_i^* = p_j^*$.

Weak scaling (for cases (i), (ii)): For F as above, $P^* = \Pi(F, Q)$ always satisfies

$$p_i^* = \frac{t}{q_i + q_j} q_i, \quad p_j^* = \frac{t}{q_i + q_j} q_j.$$

Theorem 1. *An inference rule Π is regular and local iff $\Pi(F, Q)$ is the minimizer subject to $P \in F$ of a "distance"*

$$d(P, Q) = \sum_{k=1}^{n} f_k(p_k, q_k),$$

defined by functions $f_k(p, q) \geq 0 = f_k(q, q)$, continuously differentiable in p, in cases (i),(ii) with $\frac{\partial}{\partial p} f_k(p, q) \to -\infty$ as $p \to 0$, and such that $d(P, Q)$ is strictly quasiconvex in P.

In cases (ii),(iii), these functions f_k are necessarily convex in p.

Theorem 2. *An inference rule as in Theorem 1 satisfies*

(a) transitivity iff the functions f_k are of form

$$f_k(p, q) = \varphi_k(p) - \varphi_k(q) - \varphi_k'(q)(p - q),$$

where $\Phi(P) = \sum_{k=1}^{n} \varphi_k(p_k)$ is strictly convex; then $d(P, Q)$ is the Bregman distance

$$d(P, Q) = \Phi(P) - \Phi(Q) - [\mathit{grad}\, \Phi(Q)]^T (P - Q)$$

(b) semisymmetry iff $f_1 = \cdots = f_n$

(c) weak scaling (in cases (i), (ii)) iff the functions $f_1 = \cdots = f_n$ are of form $q f(p/q)$, f is strictly convex, $f(1) = f'(1) = 0$, and $f'(x) \to -\infty$ as $x \to 0$; then $d(P, Q)$ is the f-divergence $D_f(P \| Q)$.

Bregman distances were introduced in [11]. An axiomatic characterization (in the continuous case), and hints to various applications, appear in Jones and Byrne [34]. The corresponding inference rule satisfies transitivity because it satisfies the "Pythagorean identity"

$$d(P, Q) = d(P, \Pi(F, Q)) + d(\Pi(F, Q), Q), \ \ P \in F.$$

It is not hard to see that in both cases (i) and (ii), only I-divergence is simultaneously an f-divergence and a Bregman distance.

Corollary 1. *In cases (i), (ii), regularity + locality + transitivity + weak scaling uniquely characterize the MaxEnt inference rule, with $\Pi(F, Q)$ equal to the I-projection of Q to F.*

In cases (ii), (iii), a natural desideratum is
Scale invariance: For each feasible set F, prior guess Q, and $t > 0$, $\Pi(tF, tQ) = t\Pi(F, Q)$.
In case (iii), another desideratum is translation invariance, defined analogously.

Theorem 3. *A regular, local, transitive and semisymmetric inference rule Π satisfies*
(a) translation and scale invariance (in case (iii)) iff $\Pi(F, Q)$ equals the Euclidean projection of Q to F
(b) scale invariance (in case (ii)) iff $\Pi(F, Q)$ is the minimizer of

$$d_\alpha(P, Q) = \sum_{i=1}^{n} h_\alpha(p_i, q_i)$$

subject to $P \in F$, where $\alpha \leq 1$ and

$$h_\alpha(p, q) = \begin{cases} p\log(p/q) - p + q & \alpha = 1 \\ \log(q/p) + (p/q) - 1 & \alpha = 0 \\ (q^\alpha - p^\alpha)/\alpha + q^{\alpha-1}(p - q) & \textit{else} \end{cases}$$

Remark 3. $\alpha = 1$ gives I-divergence, $\alpha = 0$ Itakura-Saito distance. An early report of success (in spectrum reconstruction) using d_α with $\alpha = 1/m$ appears in [34].

Alternate characterizations of the MaxEnt and least squares inference rules involve the intuitively appealing axiom of “product consistency” in cases (i),(ii), or “sum consistency” in case (iii). This axiom applies also in the absence of a default model. Then inference via maximizing Shannon entropy resp. minimizing Euclidean norm is arrived at, see [18] for details.

5. Discussion

After surveying various axiomatic approaches to information measures, here their scientific value is briefly addressed.

Direction (A) has an extensive literature that includes many good and many weak papers. For mathematicians, good mathematics has scientific value on its own right, the controversial issue is relevance for information theory. Note, following Shannon [52], that the justification for regarding a quantity an information measure resides in the mathematical theorems, if any, demonstrating its operational significance. This author knows of one occasion [31] when an axiomatic approach led to a new information measure of practical interest, and of another [48] when such an approach initiated research that succeeded in finding operational meanings of a previously insignificant information measure. One benefit of axiomatic work in direction (A) is the proof that new information measures with certain desirable properties do not exist. On the other hand, this research direction has developed far beyond its origins, and became a branch of the theory of functional equations. Its main results in the last 30 years are of interest primarily for specialists of that theory.

Direction (B), only briefly mentioned here, addresses a problem of major information theoretic significance. Its full solution appears far ahead, but research in this direction has already produced valuable results. In particular, many new inequalities for Shannon entropy have been discovered, starting with [60].

Direction (C) addresses the characterization of "good" inference rules, which certainly appears relevant for the theory of inference. Such characterizations involving information measures, primarily Shannon entropy and I-divergence, and secondarily Bregman distances and f-divergences, indirectly amount to characterizations of the latter. As a preferable feature, these characterizations of information measures are directly related to operational significance (for inference).

Acknowledgement

This work was partially supported by the Hungarian Research Grant OTKA T046376.

References

1. Aczél, J.; Daróczy, Z. *On Measures of Information and Their Characterizations;* Academic Press: New York, 1975.
2. Aczél, J.; Daróczy, Z. A mixed theory of information I. *RAIRO Inform. Theory* **1978**, *12*, 149–155.
3. Aczél, J.; Forte, B.; Ng, C.T. Why Shannon and Hartley entropies are "natural". *Adv. Appl. Probab.* **1974**, *6*, 131–146.
4. Ahlswede, R.; Cai, N. An interpretation of identification entropy. *IEEE Trans. Inf. Theory* **2006**, *52*, 4198–4207.
5. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. B* **1966**, *28*, 131–142.
6. Arimoto, S. Information-theoretic considerations on estimation problems. *Information and Control* **1971**, *19*, 181–194.
7. Arimoto, S. Information measures and capacity of order α for discrete memoryless channels. In *Topics in Information Theory*, Colloq. Math. Soc. J. Bolyai 16; Csiszár, I.; Elias, P., Eds.; North Holland: Amsterdam, 1977; pp. 41–52.
8. Ben-Bassat, M. f-entropies, probability of error, and feature selection. *Information and Control* **1978**, *39*, 227–242.
9. Bennett, C.; Brassard, G.; Crépeau, C.; Maurer, U. Generalized privacy amplification. *IEEE Trans. Inf. Theory* **1995**, *41*, 1915–1923.
10. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99-109.
11. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comp. Math. and Math. Phys.* **1967**, *7*, 200–217.

12. Campbell, L.L. A coding theorem and Rényi's entropy. *Information and Control* **1965**, *8*, 423–429.
13. Chaundry, T.W.; McLeod, J.B. On a functional equation. *Edinburgh Mat. Notes* **1960**, *43*, 7–8.
14. Csiszár, I. Eine informationstheorische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.* **1963**, *8*, 85–107.
15. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **1967**, *2*, 299–318.
16. Csiszár, I. A class of measures of informativity of observation channels. *Periodica Math. Hungar.* **1972**, *2*, 191–213.
17. Csiszár, I. Information measures: a critical survey. In *Trans. 7th Prague Conference on Inf. Theory, etc.*; Academia: Prague, 1977, pp. 73–86.
18. Csiszár, I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.* **1991**, *19*, 2032–2066.
19. Csiszár, I. Generalized cutoff rates and Rényi information measures. *IEEE Trans. Inf. Theory* **1995**, *41*, 26-34.
20. Daróczy, Z. Über die gemeinsame Charakterisierung der zu den nicht vollständigen Verteilungen gehörigen Entropien von Shannon und von Rényi. *Z. Wahrscheinlichkeitsth. Verw. Gebiete* **1963**, *1*, 381–388.
21. Daróczy, Z. Über Mittelwerte und Entropien vollständiger Wahrscheinlichkeitsverteilungen. *Acta Math. Acad. Sci. Hungar.* **1964**, *15*, 203–210.
22. Daróczy, Z. Generalized information functions. *Information and Control* **1970**, *16*, 36–51.
23. Daróczy, Z. On the measurable solutions of a functional equation. *Acta Math. Acad. Sci. Hungar.* **1971**, *34*, 11–14.
24. Daróczy, Z.; Maksa, Gy. Nonnegative information functions. In *Analytic Function Methods in Probability and Statistics*, Colloq. Math. Soc. J. Bolyai 21; Gyires, B., Ed.; North Holland: Amsterdam, 1979; pp. 65–76.
25. Diderrich, G. The role of boundedness in characterizing Shannon entropy. *Information and Control* **1975**, *29*, 149–161.
26. Ebanks, B.; Sahoo, P.; Sander, W. *Characterizations of Information Measures;* World Scientific: Singapore, 1998.
27. Faddeev, D.K. On the concept of entropy of a finite probability scheme (in Russian). *Uspehi Mat. Nauk* **1956**, *11*, 227–231.
28. Fischer, P. On the inequality $\sum p_i f(p_i) \geq \sum p_i f(q_i)$. *Metrika* **1972**, *18*, 199–208.
29. Forte, B. Why Shannon's entropy. In *Conv. Inform. Teor., Rome 1973*, Symposia Math. 15; Academic Press: New York, 1975; pp. 137–152.
30. Grünwald, P.; Dawid, P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Statist.* **2004**, *32*, 1367-1433.
31. Havrda, J.; Charvát, F. Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika* **1967**, *3*, 30–35.

32. Ingarden, R.S.; Urbanik, K. Information without probability. *Colloq. Math.* **1962**, *9*, 131–150.
33. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
34. Jones, L.K.; Byrne, C.L. General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis. *IEEE Trans. Inf. Theory* **1990**, *36*, 23–30.
35. Kampé de Fériet, J.; Forte, B. Information et probabilité. *C. R. Acad. Sci. Paris A* **1967**, *265*, 110–114, 142–146, and 350–353.
36. Kannappan, Pl.; Ng, C.T. Measurable solutions of functional equations related to information theory. *Proc. Amer. Math. Soc.* **1973**, *38*, 303–310.
37. Kannappan, Pl. and Ng, C.T. A functional equation and its applications in information theory. *Ann. Polon. Math.* **1974**, *30*, 105–112.
38. Kolmogorov, A.N. A new invariant for transitive dynamical systems (in Russian). *Dokl. Akad. Nauk SSSR* **1958**, *119*, 861–864.
39. Kullback, S. *Information Theory and Statistics;* Wiley: New York, 1959.
40. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79–86.
41. Lee, P.M. On the axioms of information theory. *Ann. Math. Statist.* **1964**, *35*, 415–418.
42. Linhard, J.; Nielsen, V. Studies in statistical dynamics. *Kong.Danske Vid. Selskab Mat-fys. Med.* **1971**, *38*, 9, 1-42.
43. Maksa, Gy. On the bounded solutions of a functional equation. *Acta Math. Acad. Sci. Hungar.* **1981**, *37*, 445–450.
44. Matúš, F. Infinitely many information inequalities. In *IEEE ISIT07 Nice, Symposium Proceedings;* pp. 41–44.
45. Neumann, J. Thermodynamik quantenmechanischer Gesamtheiten. *Gött. Nachr.* **1927**, *1*, 273–291.
46. Paris, J.; Vencovská, A. A note on the inevitability of maximum entropy. *Int'l J. Inexact Reasoning* **1990**, *4*, 183–223.
47. Pippenger, N. What are the laws of information theory? In *Special Problems on Communication and Computation Conference;* Palo Alto, CA, Sep. 3–5, 1986.
48. Rényi, A. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Statist. Probability, 1960;* Univ. Calif. Press: Berkeley 1961; Vol. 1, pp. 547–561.
49. Rényi, A. On the foundations of information theory. *Rev. Inst. Internat. Stat.* **1965**, *33*, 1–4.
50. Sanov, I.N. On the probability of large deviations of random variables (in Russian). *Mat. Sbornik* **1957**, *42*, 11–44.
51. Schützenberger, M.P. Contribution aux applications statistiques de la théorie de l'information. *Publ. Inst. Statist. Univ. Paris* **1954**, *3*, 3–117.
52. Shannon, C.E. A mathematical theory of communication. *Bell System Tech. J.* **1948**, *27*, 379–423 and 623–656.
53. Shore, J.E.; Johnson, R.W. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37.
54. Sibson, R. Information radius. *Z. Wahrscheinlichkeitsth. Verw. Gebiete* **1969**, *14*, 149–161.

55. Tsallis,C. Possible generalizations of the Boltzmann-Gibbs statistics. *J. Statist. Phys.* **1988**, *52*, 479-487.
56. Tverberg, H. A new derivation of the information function. *Math. Scand.* **1958**, *6*, 297–298.
57. Vajda, I. Bounds on the minimal error probability for testing a finite or countable number of hypotheses (in Russian). *Probl. Inform. Transmission* **1968**, *4*, 9–17.
58. Wald, A. *Sequential Analysis;* Wiley: New York, 1947.
59. Yeung, R.W. *A First Course in Information Theory;* Kluwer: New York, 2002.
60. Zhang, Z.; Yeung, R.W. On characterizations of entropy function via information inequalities. *IEEE Trans. Inf. Theory* **1998**, *44*, 1440–1452.