Univ. Beograd. Publ. Elektrotehn. Fak.
Ser. Mat. Fiz. № 602 — № 633 (1978), 65—72.

609. A MIXED THEORY OF INFORMATION — VI: AN EXAMPLE AT LAST: A PROPER DISCRETE ANALOGUE OF THE CONTINUOUS SHANNON MEASURE OF INFORMATION (AND ITS CHARACTERIZATION)

J. Aczél

Dedicated to D. S. Mitrinović on his 70th birthady.

1. In a series of papers [**5, 2, 6, 13, 3**] a mixed theory of information has been proposed, mainly in an axiomatic manner, where measures of information may depend both upon the events (elements of a ring of sets, messages, outcomes of experiments, weather, market situations, answers to questionnaires, subranges of values of a random variable, etc.) and their probabilities. In particular in [**5**] and [**3**], from certain hypotheses, the following general forms of these so called "inset entropies" were found

$$(1)\qquad I_n\begin{pmatrix}X_1,\dots,X_n\\ p_1,\dots,p_n\end{pmatrix}=c\sum_{i=1}^n p_i\log p_i+\sum_{i=1}^n p_i g(X_i)-g\left(\bigcup_{i=1}^n X_i\right)$$

$$(0\log 0:=0,\quad p_i\geqq 0,\quad \Sigma p_i=1,\quad X_i\cap X_j=0,\quad j\neq i=1,\dots,n)$$

and, in particular, if $\bigcup\limits_{i=1}^n X_i=\Omega$ (the certain event), then

$$(2)\qquad I_n\begin{pmatrix}X_1,\dots,X_n\\ p_1,\dots,p_n\end{pmatrix}=c\sum_{i=1}^n p_i\log p_i+\sum_{i=1}^n p_i g(X_i).$$

In both formulas, c is an arbitrary constant, g an arbitrary real valued function of events (subsets of Ω). (It will not be assumed here that the reader is familiar with the papers quoted above.)

The requirements, which have produced these formulas (those relevant for the present paper will be repeated in the Theorem below), are rather natural and it is clearly of advantage, even desirable (cf. also [**19**], where formally similar expressions have also been arrived at and applied), to allow measures of information depend also explicitly upon the events, not only their probabilities. But, requests were often made, and rightly so, for a nontrivial example of a previously known information measure which fits better into the new theory than into the traditional one.

In section 2 of this paper we give a more orthodox example than that in [**19**], this one arising directly from the heart of the classical Shannon-Wiener theory. In accordance with our personal tastes and interests, we give also a set of properties which characterize this quantity among those described in (1).

2. B. Forte has (verbally) called our attention to the paradox which has raised quite a bit of argument (see, e.g. **[23, 24, 9, 8, 25, 21, 20]**, also about other disadvantages of (3)), that the usual measure

$$-\int_u^v \rho(t)\log\rho(t)\,\mathrm{d}t \tag{3}$$

of uncertainty for continuous probability distributions (ρ is the probability frequency function) is, contrary to one's first impression, *not* the limit of the Shannon entropy for discrete distributions

$$-\sum_{i=1}^{n} p(t_i)\log p(t_i). \tag{4}$$

It is the limit of (all logarithms are of base 2)

$$-\sum_{i=1}^{n} \rho(\tau_i)\log\rho(\tau_i)\,(t_i-t_{i-1}) \qquad (\tau_i\in]t_{i-1}, t_i]),$$

that is, with appropriate choice of the $\tau_i\,(i=1, \dots, n)$ and with the distribution function $F\,(F'=\rho,\ F\,(u)=0,\ F(v)=1)$, the limit of

$$-\sum_{i=1}^{n} (F(t_i)-F(t_{i-1}))\log\frac{F(t_i)-F(t_{i-1})}{t_i-t_{i-1}}. \tag{5}$$

If, as usual, $F(t_i)-F(t_{i-1})=p_i$ is interpreted as the probability belonging to the interval $X_i=]t_{i-1}, t_i]$, $i=1, \dots, n$, $\bigcup\limits_{i=1}^{n} X_i=]u, v]=U$, then (5) goes over into an inset entropy

$$-\sum_{i=1}^{n} p_i\log p_i+\sum_{i=1}^{n} p_i\log l(X_i) \tag{6}$$

(where $l(X_i)$ is the length of X_i), clearly an expression of the form (2).

As it has also been pointed out to us, contrary to (4), the amount (6) is not necessarily nonnegative since (3) may be positive for some probability distributions and negative for others. (See, e.g., **[18]**). Our characterizations **[5, 3]** did not contain any nonnegativity supposition either. — On the other hand, for $p_j=1$, $p_i=0$ $(i\neq j)$ that is, if the value of the random variable certainly falls into X_j, then (6) reduces to

$$\log l(X_j). \tag{7}$$

In particular, if $n=1$, $X_1=U=]u, v]$, then we get

$$\log l(U) \tag{8}$$

as measure of uncertainty, if we know only that the value of the random variable falls into U, but don't know its probability distribution. If we know the distribution, the uncertainty reduces to (6). The difference between those two uncertainties is the amount

$$(9)\qquad S_n\begin{pmatrix}X_1,\dots,X_n\\ p_1,\dots,p_n\end{pmatrix}=\log l\left(\bigcup_{i=1}^{n}X_i\right)+\sum_{i=1}^{n}p_i\log p_i-\sum_{i=1}^{n}p_i\log l(X_i)$$
$$=\sum_{i=1}^{n}p_i\log\frac{p_i\,l(U)}{l(X_i)}$$

of information gained from the probability distribution.

Clearly (9) *is a* fine *example of the "inset" masures* (1) [with $c=1$, $g(X)=-\log l(X)$]. Furthermore, by SHANNON's inequality, (9) is nonnegative. Even more importantly, putting again $p_j=1$, $p_i=0$ $(i\neq j)$ into (9), we get

$$(10)\qquad S_n\begin{pmatrix}X_1,\dots,X_j,\dots,X_n\\ 0,\ \dots,1,\dots,\ 0\end{pmatrix}=-\log\frac{l(X_j)}{l(U)}$$

as the measure of information gained from the knowledge that the value of the random variable lies in the subinterval X_j of U. But (10) is exactly the measure of information introduced by N. WIENER **[26]** which, together with SHANNON's measures (3) and (4), was at the source of the whole (then) new information theory in 1948.

3. The question arises (raised also by F. ZORZITTO on a functional equations seminar at the University of Waterloo), *what characterizes this "Shannon-Wiener inset information"* (9) *among the inset measures* (1). B. FORTE has conjectured that invariance under shift and change of scale (homothecy: stretching or shrinking) with some regularity conditions would do the trick. This we can formulate (under somewhat weaker conditions) and prove in the following way.

We consider now the general formula (1) in the case where the "event X_i" stands for "the value of the random variable falls into the interval X_i". We put, again, $p_j=1$, $p_i=0$ $(i\neq j)$, this time into (1). We know by now what this means: information arising from the *knowledge* that the value of the random variable falls exactly into X_j among the subintervals. We will formulate our conditions for the quantity

$$(11)\qquad f_j(X_1,\dots,X_n)=I_n\begin{pmatrix}X_1,\dots,X_{j-1}, & X_j, & X_{j+1},\dots,X_n\\ 0,\ \dots,0, & 1, & 0,\ \dots,\ 0\end{pmatrix}$$

thus obtained, that is, see (1), for

$$(12)\qquad \begin{cases} f_j(X_1,\dots,X_n)=g(X_j)-g\left(\bigcup_{i=1}^{n}X_i\right)=g(X_j)-g(U)\\ \left(X_i=]t_{i-1},t_i]\qquad (i=1,\dots,n),\quad U=\bigcup_{i=1}^{n}X_i=]t_0,t_n]=]u,v]\right).\end{cases}$$

The shift invariance means, for (11),

$$(13)\quad f_j(X_1+s,\dots,X_n+s)=f_j(X_1,\dots,X_n)\ (X_i+s=]t_{i-1}+s,t_i+s],\ i=1,\dots,n)$$

and, with (12),

$$(14)\quad g(x+s,y+s)-g(u+s,v+s)=g(x,y)-g(u,v)\qquad (u\leqq x<y\leqq v\in\mathbf{R},\ s\in\mathbf{R}).$$

Here we wrote, for the sake of simplicity, $t_{j-1}=x$, $t_j=y$ and

$$g(x, y):=g(]x, y])=g(X_j). \tag{15}$$

We choose $s=-x=-u$, $(x=x_{j-1}=x_0=u$, thus $j=1)$ and $v=b$ (constant, choose it large and/or extend later) in (14) and get $g(x, y)=g(0, y-x)+g(x, b)-g(0, b-x)$, that is,

$$g(x, y)=\alpha(y-x)+\beta(x). \tag{16}$$

(for all $x<y\leqq b$ but, since we could choose b as large as we wanted, (16) holds for all $x<y$), where

$$\alpha(z)=g(0, z) \tag{17}$$

and

$$\beta(x)=g(x, b)-g(0, b-x). \tag{18}$$

Putting (16) back into (14), we have $\beta(x+s)-\beta(u+s)=\beta(x)-\beta(u)$ or, by holding u constant (small), $\beta(x+s)=\beta(x)+\gamma(s)$.

This is a Pexider equation **[1]**. If g is measurable, so is β, by (18), and thus $\beta(x)=Bx+C$, $(B, C$ constants) and, cf. (16),

$$g(x, y)=\alpha(y-x)+Bx \tag{19}$$

(we have sumberged C into α).

Invariance against homothecy (change of scale) means, on the other hand,

$$f_j(X_1t, \ldots, X_nt)=f_j(X_1, \ldots, X_n) \quad (t>0,\ X_it=]t_{i-1}t, t_it],\ i=1, \ldots, n) \tag{20}$$

that is, cf. (12), (15),

$$g(xt, yt)-g(ut, vt)=g(x, y)-g(u, v) \qquad (u\leqq x<y\leqq v,\ t>0). \tag{21}$$

Substituting (19) into (21), we get

$$\alpha[t(y-x)]+Btx-\alpha[t(v-u)]-Btu=\alpha(y-x)+Bx-\alpha(v-u)-Bu. \tag{22}$$

If we choose $y=x+d$ (d constant), the comparison of the coefficients of x on both sides of (22) gives $B=0$, reducing (19) to

$$g(x, y)=\alpha(y-x) \qquad (x<y) \tag{23}$$

and (22), with $z=y-x>0$, $w=v-u>0$, to

$$\alpha(tz)-\alpha(tw)=\alpha(z)-\alpha(w) \qquad (z>0,\ w>0,\ t>0)$$

or, holding z constant,

$$\alpha(tw)=\alpha(w)+\delta(t) \quad \text{for all} \quad t, w>0.$$

This is again a Pexider equation and since, if g is measurable, so is α, [cf. (17)], we have $\alpha(t)=a\log t+b$

Thus we get from (23), finally,

$$g(x, y) = a\log(y-x)+b \qquad (x<y;\quad a,\, b\quad \text{constants}). \tag{24}$$

Substitution of (15) and (24) into (1) leaves us with

$$\left\{\begin{aligned} &I_n\begin{pmatrix} X_1, \dots, X_n \\ p_1, \dots, p_n \end{pmatrix} = c\sum_{i=1}^{n} p_i \log p_i + a\sum_{i=1}^{n} p_i \log l(X_i) - a\log l\left(\bigcup_{i=1}^{n} X_i\right), \\ &\quad (p_i \geqq 0,\ \ \Sigma p_i = 1,\ \ X_i =]\,t_{i-1},\ t_i],\ \ t_i > t_{i-1},\ \ l(X_i) = t_i - t_{i-1}, \\ &\qquad\qquad i = 1, \dots, n;\quad a,\, c\ \text{arbitrary constants}), \end{aligned}\right. \tag{25}$$

very similar to (9). In view of [5], we have (almost) proved the following.

Theorem. *The information, gained from the knowledge of the finite (discrete) probability distribution of the values of a random variable on the straight line, is given by* (25) *if and only if the following conditions are satisfied.*

(i) I_n *is symetric in columns.*

(ii) I_n *is recursive, that is*

$$I_n\begin{pmatrix} X_1, X_2, X_3, \dots, X_n \\ p_1, p_2, p_3, \dots, p_n \end{pmatrix} = I_{n-1}\begin{pmatrix} X_1 \cup X_2, X_3, \dots, X_n \\ p_1+p_2, p_3, \dots, p_n \end{pmatrix} + (p_1+p_2)\, I_2\begin{pmatrix} X_1 & , & X_2 \\ \dfrac{p_1}{p_1+p_2} & , & \dfrac{p_2}{p_1+p_2} \end{pmatrix},$$

$$\left(n>2;\quad 0\cdot I_2\begin{pmatrix} X_1, X_2 \\ 0/0, 0/0 \end{pmatrix} = 0\right).$$

(iii) *The functions* $t \mapsto I_2\begin{pmatrix} X_1, X_2 \\ 1-t, t \end{pmatrix}$ *and*

$$(t_{j-1},\ t_j) \mapsto I_j\begin{pmatrix} X_1, \dots, X_{j-1}, X_j, X_{j+1}, \dots, X_n \\ 0, \dots, 0\ \ , 1,\ 0,\ \ \dots, 0 \end{pmatrix} \quad (X_j =]\,t_{j-1},\ t_j]), \tag{26}$$

are measurable (for one j*).*

(iv) *The functions* I *or just* f_j *(for one* j*), as defined in* (11), *are invariant under shift and change of scales, that is,* (13) *and* (20) *are satisfied.*

Proof. Only a few additions are needed to the above proof. In [5], the result (1) has been proved from (i), (ii) and the first part of (iii). There is just one hitch. In [5] also $X_j = \varnothing$ (the empty set) was permitted, while here $t_{j-1} < t_j$ was supposed. However, the proof in [5] has been carefully carried out that way that it remains valid if $X_j = \varnothing$ implies $p_j = 0$. Then, of course, there is no problem in extending the definition of our I_n and the validity of (1) and of the proofs in this paper, to the case when some $t_j - t_{j-1} = 0 = p_j$ [since they will not change the right hand side of (1)].

Notice also that we needed the conditions (13) and (20) only for one j. (At one point we chose $j=1$ because of $x_{j-1} = u$, but this is possible now for any j, in view of the extension which we have just made).

Finally, as partly mentioned above, if the function defined in (26) is measurable, so are, by (12), (15), (17), and (18), α and β.

Thus we have established everything needed for the argument preceeding the Theorem, which is now proved.

4. However, (25) is still not (9). In order to characterize (9), we consider the possibility that

$$p_i = \frac{l(X_i)}{l(U)} \qquad (i=1, \ldots, n) \tag{27}$$

that is, the distribution is uniform, the probability just equals the quotient of lengths of the respective subinterval and of the whole interval. Then the quantity (9) *reduces to zero* as it might well do: we did not really gain any new information by knowing *this* probability distribution. We can see this also in the following way. If we put (27) into (6), we get exactly the amount (8) obtained *without* knowledge of any probability distribution. If this condition (*no new information is gained from the knowledge that the probability distribution is uniform on the subinterval when, from all we knew, it was uniform on the whole interval*)

$$\text{(v)} \qquad I_n\begin{pmatrix} X_1, \ldots, X_n \\ l(X_1)/l(U), \ldots, l(X_n)/l(U) \end{pmatrix} = 0$$

is added to those which have given us (25), we get $a = -c$, that is, cf. (9),

$$I_n\begin{pmatrix} X_1, \ldots, X_n \\ p_1, \ldots, p_n \end{pmatrix} = cS_n\begin{pmatrix} X_1, \ldots, X_n \\ p_1, \ldots, p_n \end{pmatrix}. \tag{28}$$

Clearly already

$$\text{(vi)} \qquad I_2\begin{pmatrix} X_1, X_2 \\ \frac{1}{2}, \frac{1}{2} \end{pmatrix} = 0 \quad \text{if} \quad l(X_1) = l(X_2) \left(= \frac{1}{2} l(U)\right)$$

does the trick, giving us $a = -c$ and thus (28).

Finally, if we insist on obtaining exactly S_n, without even the multiplicative constant c, we just have to choose the unit of information appropriately (one bit), for instance by postulating

$$\text{(vii)} \qquad f_1(X_1, X_2) = I_2\begin{pmatrix} X_1, X_2 \\ 1, 0 \end{pmatrix} = 1 \quad \text{if} \quad l(X_1) = l(X_2)$$

that is, *we get one bit of information from the knowledge that the value of a random variable falls into a given half of the whole interval.* Thus we have proved the following:

Corollary. *If, in addition to* (i) – (iv), *we have also* (v), *or even just* (vi), *then we get* (28) [*cf.* (9)] *and if we have* (vii) *on top of all, then* I_n *is the "Shannon-Wiener inset entropy"* (9).

REMARK. By writing $q_i = \frac{l(X_i)}{l(U)}$, the "SHANNON-WIENER inset entropy" (9) [but not (6)] goes over into a (purely) probabilistic directed divergence

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

for which several characterizations are known (see, e.g. **[15, 11, 14, 12, 4, 10, 13]**˙ But we did not need them there.

Indeed, while the controversies over the entropies (3) and (4), mentioned at the beginning, and the advantages of (9) have lead to the preference of directed divergences over entropies (in particular for continuous distributions, cf., also for interpretations, [**23, 9, 24, 25**]), the present paper shows that the key quantities (6) and (9) are special cases of "inset" *entropies.* So are also the somewhat more general expresions in [**23, 24, 9, 25**] (cf. [**22, 16, 17, 7, 10**]). — We see also that the union of events $U =]u, v]$ may vary here from case to case. So it is of advantage that we took in (1) conditional probabilities, *relative to* the union of events and that we did not restrict ourselves to the case where this union is universally fixed (Ω).

This research has been supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. J. Aczél: *Lectures on Functional Equations and their Applications.* New York—London 1966.
2. J. Aczél: *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. J. Aczél: *A mixed theory of information* — V: *How to keep the (inset) expert honest.* Selecta Statistica Canadiana **4** (1979).
4. J. Aczél — Z. Daróczy: *On Measures of Information and their Characterizations.* New York—San Francisco—London (1975).
5. J. Aczél — Z. Daróczy: *A mixed theory of information* — I: *Symmetric recursive and measurable entropies of randomized systems of events.* Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Informat. Théor. **12** (1978), 149—155.
6. J. Aczél — Pl. Kannappan: *A mixed theory of information* — *III: Inset entropies of degree ß.* Information and Control **39** (1978).
7. I. Csiszár: *On generalized entropy.* Studia Sci. Math. Hungar. **4** (1969), 401—419.
8. B. Forte: *Il principio di minima certezza e le sue prime applicazioni alla meccanica statistica,* Ann. Univ. Ferrara Sez. VII. **12** (1966), 7—17.
9. E. T. Jaynes: *Information theory and statistical mechanics.* Statistical Physics (Brandeis Summer Institute 1962, vol. 3), W. A. Benjamin, New York, 1963, 181—218.
10. R. W. Johnson: *Axiomatic characterization of a family of information measures that contains the directed divergence.* Naval Research Laboratory Memorandum Report **3646**, Washington 1977.
11. Pl. Kannappan: *On directed divergence and inaccuracy.* Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **25** (1972), 49—55.
12. Pl. Kannappan: *On Rényi-Shannon entropy and related measures.* Selecta Statistica Canadiana **2** (1974), 65—76.
13. Pl. Kannappan: *A mixed theory of information* — IV: *Inset inaccuracy and directed divergence.* Metrika (1979).
14. Pl. Kannappan — C. T. Ng: *Measurable solutions of functional equations related to information theory.* Proc. Amer. Math. Soc. **38** (1973), 303—310.
15. Pl. Kannappan — P. N. Rathie: *On various characterizations of directed divergence.* Trans. 6th Prague Conf. Information Theory, Statist. Decision Functions, Random Proc. 1971, Academia, Prague 1973, 331—339.
16. S. Kullback: *Information Theory and Statistics.* New York, London 1959.
17. S. Kullback — R. A. Leibler: *On information and sufficiency.* Ann. Math. Statist **22** (1951), 79—86.
18. E. H. Lieb: *Some convexity and subadditivity properties of entropy.* Bull. Amer. Math. Soc. **81** (1975), 1—13.

19. J. R. Meginnis: *A new class of symmetric utility rules for gambles, subjective marginal probability functions, and a generalized Bayes rule.* Business and Economic Stat. Sec. Proc. Amer. Stat. Assoc. 1976, 471—476.

20. M. Mugur-Schächter — C. Padet — J. P. Padet: *Le concept de quantité d'information accessible par mesure correspondant à une distribution de probabilité continue.* C. R. Acad. Sci. Paris **282** (1976), A487—A490.

21. C. Padet — M. Mugur-Schächter — J. P. Padet: *Une nouvelle expression de l'entropie (ou information moyenne) pour les distributions statistiques continues.* C. R. Acad. Sci. Paris **281** (1975), A993—A996.

22. A. Pérez: *Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales.* Trans. 1st Prague Conf. Information Theory, Statist. Decision Functions, Random Proc. 1956, Czech. Acad. Sci., Prague, 1957, 183—208.

23. I. Vincze: *An interpretation of the I-divergence of information theory.* Trans. 2nd Prague Conf. Information Theory, Statist. Decision Functions, Random Proc. 1959, Czech. Acad. Sci., Prague, 1960, New York—London, 1961, 681—684.

24. I. Vincze: *Some questions on the probabilistic concept of information* (Hung.) Magyar Tud. Akad. Mat. Fiz. Oszt. Közl. **12** (1962) 7—14 (English Transl. in Selected Translations in Mathematical Statistics and Probability **5** (1965), 373—380).

25. I. Vincze: *On the maximum probability principle.* Progress in Statistics (9th European Meeting of Statisticians, Budapest 1972). Vol. 2, Amsterdam—London; J. Bolyai Math Soc., Budapest, 1974, 260—393.

26. N. Wiener: *Cybernetics, or Control and Communication in the Animal and the Machine.* Paris, Cambridge, New York 1948.

Fac. of Math. University of Waterloo
Waterloo, Ontario, Canada N2L 3G1