ON THE CHARACTERIZATION OF SHANNON'S ENTROPY BY SHANNON'S INEQUALITY

Dedicated to the memory of Hanna Neumann

J. ACZÉL and A. M. OSTROWSKI

(Received 15 May 1972, revised 9 January 1973)

Communicated by J. B. Miller

1. In [2, 5, 6, 7] a.o. several interpretations of the inequality

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

for all

$$p_k > 0,\ q_k > 0\ (k = 1, 2, \cdots, n) \text{ such that } \sum_{k=1}^{n} p_k = \sum_{k=1}^{n} q_k = 1 \tag{2}$$

were given and the following was proved.

If the inequality (1) is satisfied *for a fixed n greater than two* on the domain (2) and *if f is differentiable on the open interval* $]0, 1[$, then and only then there exist two constants $a \geqq 0$, b so that

$$f(p) = a \log p + b \quad \text{for all} \quad p \in]0, 1[. \tag{3}$$

We mention here only two interpretations. The first is the following. We ask from an expert his estimations on a certain probability distribution (outcomes of an experiment, market situation, weather, etc.). He gives this as $(q_1, q_2, \cdots, q_n)$ while his subjective probabilities for the same events are $(p_1, p_2, \cdots, p_n)$. Suppose that he agrees to be paid only after the outcome of the experiment (market situation, etc.) is known and that his payoff will be $f(q_k)$ if the k-th event happens. Then his expected earning will be

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

In order to "keep the expert honest" it seems wise (for the customer) to choose the "payoff function" f so that the expert's expected earning will be maximal if

he has given his subjective probabilities as estimates for the customer, i.e. so that on the domain (2) the inequality (1)

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

holds.

The other interpretation is connected with Shannon's inequality

$$(4) \qquad -\sum_{k=1}^{n} p_k \log q_k \geqq -\sum_{k=1}^{n} p_k \log p_k$$

on (2) which is rather important in coding theory [see e.g. 3]. The quantity on the right is Shannon's entropy. One can ask which functions g satisfy, like in (4) the negative logarithm, an inequality of the form

$$(5) \qquad \sum_{k=1}^{n} p_k g(q_k) \geqq \sum_{k=1}^{n} p_k g(p_k)$$

on (2).

Evidently, if a function g satisfies (5), then $f = -g$ satisfies (1) and vice versa, so the general (differentiable) solution of (5) on (2) for a fixed $n > 2$ is given on $]0, 1[$ by

$$(6) \qquad g(p) = -a \log p + b \qquad (a \geqq 0, b \text{ arbitrary constants}).$$

With these g, the right-hand side of (5) still is the Shannon entropy up to a (non-negative) multiplicative and an additive constant. So the above and the more general theorem to be proved in this note are also characterizations of the Shannon entropy.

We choose here to formulate the results in terms of (1) rather than (5). The implications on (5) are obvious.

It has been conjectured in [1] and proved by Fischer in [4] that the condition of differentiability can be discarded in the above result. Rényi has written but not published a modified version of this proof. Since we think that Rényi's elegant proof should be published to which his early death gives tragic actuality, and since we have succeeded to further shorten and simplify his proof even in two different ways and also to generalize it slightly (Rényi has supposed (1) for a fixed $n > 2$ *and for* $n = 2$, we do not need the latter), we give here these modified proofs. We mention yet, that the same theorem was announced in [8] with credit given to Gleason, but without proof and without the restriction $n > 2$ (without which it is not true, a counter example being $f(p) = 2p - p^2$, [for detailed discussions of the case $n = 2$ see 2, 4, 9]). However, Gleason has sent (later than Fischer and Rényi, but independently) a correct proof of the same theorem to one of the authors of the

present paper. His proof was in many respects similar to that of Fischer and Rényi, but longer.

2. We give now our two versions of the proof (they differ only in a few steps).

THEOREM. *If f satisfies the inequality*

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

for one $n > 2$ and for all $p_1, p_2, \cdots, p_n, q_1, q_2, \cdots, q_n$ such that

$$p_k > 0,\ q_k > 0\ (k = 1, 2, \cdots, n),\ \sum_{k=1}^{n} p_k = \sum_{k=1}^{n} q_k = 1, \tag{2}$$

then and only then there exist constants $a \geqq 0$ and b so that

$$f(p) = a \log p + b \text{ for all } p \in]0, 1[. \tag{3}$$

PROOF. We get by multiplying (4) by $(-a)$ and by adding b that (3) (with $a \geqq 0$) satisfies (1) on (2). (The Shannon inequality (4) is a well-known consequence of the inequality between the arithmetic and the geometric means: [3]). Our two ways of proving that the validity of (1) on (2) implies (3) have their first steps in common:

(i) f is nondecreasing,

while the second steps are different in the two proofs:

(iia) all Dini derivatives are equal in every point and

$$pf'(p) = a \text{ (constant)} \quad a \geqq 0 \quad (p \in]0, 1[) \tag{7}$$

resp.

(iib) the function $p \mapsto pD_+ f(p)$ is a nonnegative finite constant on $]0, 1[$.

PROOF OF (i). Put into (1) $p_k = q_k$ $(k \geqq 3)$ in order to get

$$p_1[f(q_1) - f(p_1)] \leqq p_2[f(p_2) - f(q_2)] \tag{8}$$

for all p_1, p_2, q_1, q_2 satisfying

$$p_1 > 0,\ p_2 > 0,\ q_1 > 0,\ q_2 > 0,\ p_1 + p_2 = q_1 + q_2 < 1. \tag{9}$$

The conditions in (9) remain unchanged if we interchange the pair (p_1, p_2) with (q_2, q_1), so the inequality (8) remains true also if we write p_1 and p_2 instead of q_2 and q_1 and vice versa:

$$q_2[f(p_2) - f(q_2)] \leqq q_1[f(q_1) - f(p_1)]. \tag{10}$$

Multiply (8) by q_2 and (10) by p_2 and compare the two inequalities. We get

$$q_2 p_1[f(q_1) - f(p_1)] \leqq p_2 q_1[f(q_1) - f(p_1)] \tag{11}$$

for all p_1, p_2, q_1, q_2 satisfying (9).

If $p_1 < q_1$, then (9) implies $q_2 < p_2$, so (11) will hold iff $f(q_1) - f(p_1) \geqq 0$. Thus

$$f(p_1) \leqq f(q_1) \text{ if } p_1 < q_1, \tag{12}$$

that is, f is nondecreasing. The inequality (12) holds for all p_1, q in $]0,1[$, because then p_2, q_2 can be found so that (9) be satisfied. Thus f is *nondecreasing in* $]0,1[$ and (i) is proved.

We will need two consequences of (8).

Put $q_1 = p_1 + \delta$, $q_2 = p_2 - \delta$ into (8) and (9), in order to get, after division by δ,

$$p_1 \frac{f(p_1 + \delta) - f(p_1)}{\delta} \leqq p_2 \frac{f(p_2) - f(p_2 - \delta)}{\delta} \tag{13}$$

for all p_1, p_2, δ satisfying $0 < \delta < p_2$ and

$$p_1 + p_2 < 1, \ p_1 > 0, \ p_2 > 0. \tag{14}$$

Now put into (8) (and (9)) $p_1 = q_1 + \delta$, $p_2 = q_2 - \delta$ in order to get, after division by $(-\delta)$,

$$(q_2 - \delta) \frac{f(q_2) - f(q_2 - \delta)}{\delta} \leqq (q_1 + \delta) \frac{f(q_1 + \delta) - f(q_1)}{\delta} \tag{15}$$

for all q_1, q_2, δ satisfying $0 < \delta < q_2$ and

$$q_1 + q_2 < 1, \ q_1 > 0, \ q_2 > 0. \tag{16}$$

3. PROOF OF (iia) AND FIRST PROOF OF THE THEOREM. The function f, being monotonic, is differentiable almost everywhere in $]0,1[$. Fix a point $r \in]0, \varepsilon[$ $(0 < \varepsilon < 1)$ at which f is differentiable. We will prove that f' exists and (7) holds for every $p \in]0, 1 - \varepsilon[$. Since ε can be chosen as small as we wish, this will prove (iia) for all $p \in]0,1[$. But, for the time being, we have

$$p + r < 1, \ p > 0, \ r > 0. \tag{17}$$

If we take $p_1 = p$, $p_2 = r$ (the inequalities (17) assure that (14) is satisfied) and let δ tend to 0 in such a manner that the lefthand side of (13) tend to its lim sup, then we have

$$pD^+f(p) \leqq rf'(r), \tag{18}$$

since f is differentiable at r. $(D^+, D^-, D_+, D_-$ denote the right upper, left upper right lower, left lower Dini derivatives, respectively.) If, on the other hand, we choose in (13) $p_1 = r$, $p_2 = p$ and let δ tend to 0 so that the right-hand side tend to its lim inf, i.e. to $pD_-f(p)$ then we get similarly

$$rf'(r) \leqq pD_-f(p). \tag{19}$$

Exactly the same manoeuvres as above, with (q_2, q_1) instead of (p_1, p_2), lead from (15) to

$$pD^-f(p) \leqq rf'(r) \tag{20}$$

and to

$$rf'(r) \leqq pD_+f(p). \tag{21}$$

By combining (21) with (18), and (19) with (20) since by definition $D_+ \leqq D^+$, $D_- \leqq D^-$, we have

$$rf'(r) \leqq pD_+f(p) \leqq pD^+f(p) \leqq rf'(r)$$

and

$$rf'(r) \leqq pD_-f(p) \leqq pD^-f(p) \leqq rf'(r).$$

Taking into consideration that r was fixed, so $rf'(r) = a$ (constant, nonnegative since f is nondecreasing), we have proved (iia),

$$D_+f(p) = D^+f(p) = D_-f(p) = D^-f(p) = \frac{a}{p},$$

that is, f is *everywhere differentiable* and we have (7)

$$f'(p) = \frac{a}{p} \tag{22}$$

for all $p \in]0, 1 - \varepsilon[$ and, since ε is as small as we wish, for all $p \in]0, 1[$. Equation (22) implies (3) which concludes the first proof of the Theorem.

4. The second proof does not depend on the fact that every monotonic function is almost everywhere differentiable and it does not use any other result in measure theory either. Instead it applies a more elementary theorem of Scheeffer [10]. The proof proceeds to (8), (9), (13), (15) and to the nondecreasing monotonicity of f as above and then continues in the following way.

PROOF OF (iib) AND SECOND PROOF OF THE THEOREM.

Let $\delta \searrow 0$ in (13) in such a manner that the right hand side tend to its lim inf, i.e. to $p_2D_-f(p_2)$. No cluster point of the left handside is smaller than its lim inf, that is, than $p_1D_+f(p_1)$. So we have

$$p_1D_+(p_1) \leqq p_2D_-f(p_2) \tag{23}$$

for all p_1, p_2 satisfying (14). Similarly, from (15) we get

$$q_2D_-f(q_2) \leqq q_1D_+f(q_1) \tag{24}$$

for all q_1, q_2 satisfying (16). Comparing (16) with (14) we see that (24) remains true if we replace q_1 by p_1 and q_2 by p_2. So we have

$$p_2D_-f(p_2) \leqq p_1D_+f(p_1),$$

which, together with (23), gives

$$p_1 D_+ f(p_1) = p_2 D_- f(p_2) \tag{25}$$

for all p_1, p_2 satisfying (14).

Fix now $p_2 \in]0, \varepsilon[$, then, f being nondecreasing, (14) and (25) give for arbitrary $p = p_1 \in]0, 1 - \varepsilon[$

$$pD_+ f(p) = a \geqq 0 \quad \text{(constant)} \tag{26}$$

on $]0, 1 - \varepsilon[$ and, since ε is as small as we wish, also on $]0, 1[$.

A priori a could be infinite. But then we would have from (25) and (26)

$$D_+ f(p) = \infty = D_- f(p) \text{ on }]0, 1[\tag{27}$$

and, even for arbitrarily large constants A,

$$D_+[f(p) - Ap] = D_+ f(p) - A = \infty = D_- f(p) - A = D_-[f(p) - Ap] \text{ on }]0, 1[, \tag{28}$$

in particular $p \mapsto f(p) - Ap$ would be increasing on $]0, 1[$.

On the other hand, for all $A > 2f(3/4) - 2f(1/4)$ we have

$$f\left(\frac{1}{4}\right) - A\frac{1}{4} > f\left(\frac{3}{4}\right) - A\frac{3}{4}$$

which is impossible if $p \mapsto f(p) - Ap$ is increasing on $]0, 1[$. Thus (27) leads to a contradiction and a in (26) is a *finite* constant, which concludes the proof of (iib).

Since $D_+ f(p)$ is finite everywhere on $]0, 1[$, the same follows by (25) for $D_- f(p)$. But then f must be *continuous* on $]0, 1[$ since a discontinuity of a monotonic function is always a jump and there either $D_+ f(p)$ of $D_- f(p)$ would be ∞.

Further we have from (26)

$$D_+ f(p) = D_+(a \log p) \text{ on }]0, 1[. \tag{29}$$

However, L. Scheeffer [10] has proved in a very elementary manner that the continuity of f and F and the validity of

$$D_+ f(p) = D_+ F(p)$$

(both finite) on an interval implies that there exists a constant b such that on this interval

$$f(p) = F(p) + b.$$

So (29) implies (3) and our Theorem is proved again.

We have used above (after (28)) the fact that a *function g is increasing on an (open) interval I, if both*

$$D_+ g(x) > 0 \text{ and } D_- g(x) > 0 \text{ for all } x \in I.$$

For completeness sake we give here a proof of this proposition. If, for $x_1 \in I$, we have

$$k = D_+ g(x_1) > 0$$

then there exists a δ such that

$$(30)\quad \frac{g(x_1+h)-g(x_1)}{h} > \frac{k}{2} > 0, \text{ i.e. } g(x_1+h) > g(x_1), \text{ whenever } 0 < h < \delta.$$

Similarly, $D_-g(x_1) > 0$ implies

$$(31)\quad g(x_1-h) < g(x_1) \text{ whenever } 0 < h < \delta.$$

We have to prove that *for any* $x_0 \in I$

$$(32)\quad g(x) > g(x_0)$$

whenever $x > x_0$ $(x \in I)$. Let x_1 be the smallest number with the property that (32) holds for all $x \in]x_0, x_1[\subseteq I$. We prove that x_1 has to be the right extremity of I. For else we had, by (30) and (31), for sufficiently small positive h,

$$g(x_0) < g(x_1-h) < g(x_1) < g(x_1+h)$$

contrary to the definition of x_1. This concludes the proof of the above proposition.

We are grateful to Professor W. Walter for a comment which has helped us to shorten the second version of the proof.

References

[1] J. Aczél, 'Problem 3, P21', *Aequations Math.* 1(1968), 300; 2 (1968), 111.

[2] J. Aczél — J. Pfanzagl, 'Remarks on the Measurement of Subjective Probability and Information', *Metrika* 11 (1966), 91–105.

[3] A. Feinstein, *Foundations of Information Theory* (McGraw Hill, New York — Toronto — London, 1958).

[4] P. Fischer, 'On the Inequality $\Sigma p_i f(p_i) \geq \Sigma p_i f(q_i)$', *Metrika* 18 (1972), 199–208.

[5] I. J. Good, 'Rational Decisions', *J. Roy. Statist. Soc.* Ser. B 14 (1952), 107–114.

[6] I. J. Good, *Uncertainty and Business Decisions* (Liverpool University Press, Liverpool, 1954, 1957), in part. p. 31 (2nd ed. 1957, p. 33).

[7] J. Marschak, 'Remarks on the Economics of Information', in *Contributions to Scientific Research and Management* (Univ. California Press, Los Angeles, Calif. 1960), pp. 79–98.

[8] J. McCarthy, 'Measures of the Value of Information', *Proc. Nat. Acad. Sci. U. S. A.* 42 (1956), 654–655.

[9] Gy. Muszély, 'On Continuous Solutions of a Functional Inequality', *Metrika* 19 (1973), 65–69.

[10] L. Scheeffer, 'Zur Theorie der stetigen Funktionen einer reellen Veränderlichen,' *Acta Math.* 5 (1884), 183–194, 279–296, in part. pp. 183–185, 279–280; contained also, e.g., in E. W. Hobson, *The Theory of Functions of a Real Variable and the Theory of Fourier's Series* (Cambridge University Press, Cambridge, 1907), in part. pp. 273–274 (3rd ed. 1927, p. 366).

University of Waterloo
Waterloo, Ont., Canada
and
Universität Basel
Basle, Switzerland