Department of Applied Analysis
and Computer Science

Research Report AA-73-05
January 1973

1. ON SHANNON'S INEQUALITY, OPTIMAL CODING, AND CHARACTERIZATIONS OF SHANNON'S AND RÉNYI'S ENTROPIES

2. DETERMINATION OF ALL ADDITIVE QUASIARITHMETIC MEAN CODEWORD LENGTHS

by

J. Aczél
Univeristy of Waterloo

The first paper was given at the Meeting on Theoretical Information of the Istituto Nazionale di Alta Matematica, Rome. The problem which was still unsolved on pp.31-32 of the first paper is solved in the second paper, to be published in the Zeitschrift für Wahrscheinlichkeitsrechnung und verwandte Gebiete.

1. On Shannon's Inequaltiy, Optimal Coding, and Characterizations of Shannon's and Rényi's Entropies

J. Aczél

J. ACZÉL: On Shannon's Inequality, Optimal Coding and Characterizations of Shannon's and Rényi's Entropies

1. The classical form of Shannon's inequality is

$$(1) \qquad -\sum_{k=1}^{N} p_k \log p_k \leq -\sum_{k=1}^{N} p_k \log q_k ,$$

for all $N \geq 2$ if

$$(2) \qquad \sum_{k=1}^{N} p_k = \sum_{k=1}^{N} q_k = 1 ; \; p_k > 0, \; q_k > 0; \; k = 1,2,\ldots,N,$$

with equality in (1) iff

$$(3) \qquad p_k = q_k \quad (k = 1,2,\ldots N).$$

The expression on the left hand side of (1) is the Shannon entropy, if we take 2 as base of the logarithms in (1).

The most important application of Shannon's inequality may be the theorem asserting that the average length of a codeword in a uniquely decipherable code cannot be smaller than the Shannon entropy divided by the logarithm (base 2) of the number of symbols in the code. So, even optimal coding cannot produce shorter average lengths of codewords, but should try to approximate this lower bound as closely as possible.

In this paper, by examining two proofs of Shannon's inequality closer, we will first extend it to situations more general than (2). This, in its turn, will simplify the proof of the optimal coding theorem, mentioned above. Also, the analysis of the equality cases of this extended inequality will give guidance, how such optimal codes can be selected.

The relevance of Shannon's inequality to Shannon's entropy is a.o. due to the fact that the quantity on the left hand side of (1) is Shannon's entropy. Conversely, by putting an unknown function in place of log in (1), we can obtain from Shannon's inequality a characterization of Shannon's entropy. We will give two proofs (really shortcuts of previous proofs) for this characterization.

Another way of characterizing Shannon's entropy is the optimal coding theorem itself. A one parameter class of entropies, the so called Rényi entropies, can be similarly characterized. We will give a new version of this characterization, in particular for positive values of the parameter, thus characterizing the Rényi entropies of positive order.

2. One proof of Shannon's inequality (cf. A. Feinstein 1958) is based on the inequality of the geometric and arithmetic means. This asserts that

$$\prod_{k=1}^{N} x_k^{p_k} \leq \sum_{k=1}^{N} p_k x_k \tag{4}$$

if

$$\sum_{k=1}^{N} p_k = 1,\ p_k > 0,\ x_k \geq 0 \quad (k = 1,2,\ldots,N). \tag{5}$$

Inequality holds in (4) iff

$$x_1 = x_2 = \ldots = x_N. \tag{6}$$

In order to prove the Shannon inequality we put into (4) $x_k = q_k/p_k$ ($k = 1,2,\ldots,N$, the conditions (5) are satisfied because of (2)), and get

$$\prod_{k=1}^{N} \left(\frac{q_k}{p_k}\right)^{p_k} \leq \sum_{k=1}^{N} p_k \frac{q_k}{p_k} = \sum_{k=1}^{N} q_k = 1 \tag{7}$$

and, taking logarithms on both sides of (7), which we can do since $p_k > 0$, $q_k > 0$ $(k = 1,2,\ldots,N)$, we get

$$\sum_{k=1}^{N} p_k (\log q_k - \log p_k) \leq 0 \tag{8}$$

from which the desired inequality (1) follows at once. There is equality in (7), cf. (6), iff

$$\frac{q_1}{p_1} = \frac{q_2}{p_2} = \ldots = \frac{q_N}{p_N} = c \ , \tag{9}$$

but, because of (2),

$$\sum_{k=1}^{N} p_k = 1 = \sum_{k=1}^{N} q_k = c \sum_{k=1}^{N} p_k \tag{10}$$

thus $c = 1$ in (9) and we have equality in (1) iff (3) holds.

When we look carefully at this proof, we see that it can be modified so that instead of

$$\sum_{k=1}^{N} q_k = 1 \tag{11}$$

we may suppose only

$$\sum_{k=1}^{n} q_k \leq 1 \ . \tag{12}$$

Indeed, then (7) will change into

$$\prod_{k=1}^{N} \left(\frac{q_k}{p_k}\right)^{p_k} \leq \sum_{k=1}^{N} p_k \frac{q_k}{p_k} = \sum_{k=1}^{N} q_k \leq 1 \tag{13}$$

and (8), (1) can still be derived. There will be equality in (1) iff there is equality in both inequalities of (13), that is, iff (9) and (11) hold, so we get again (10) and (3).

Of course, the inequality of the arithmetic and geometric means expresses the concavity of the logarithm function on $]0,\infty[$. Indeed, if

$$(14) \qquad \sum_{k=1}^{N} p_k = 1, \; p_k > 0, \; x_k > 0 \quad (k = 1,2,\dots,N),$$

then (4) is equivalent to the Jensen inequality (see e.g. G. H. Hardy - J. E. Littlewood - G. Pólya 1934, Section 3.8)

$$(15) \qquad \sum_{k=1}^{N} p_k \, \psi(x_k) \leq \psi\Big(\sum_{k=1}^{N} p_k x_k\Big)$$

for the function $\psi = \log$, with equality again exactly if (6) holds. (If the function value $-\infty$ is admissible, then we can take (5) as domain instead of (14), that is, allow some x_k to be 0.)

3. The Shannon inequality can also be obtained (see e.g. J. Aczél - Z. Daróczy 1975, cf. F. M. Reza 1961) from the concavity on $[0,1]$ of the function L defined by

$$(16) \qquad L(x) = \begin{cases} -x \log x & \text{if } x \in \,]0,1] \\ 0 & \text{if } x = 0 \end{cases}.$$

(In information theory one usually takes 2 as basis of logarithms, but this is not important here as long as the basis is greater than 1). The function L is indeed concave on $[0,1]$ since (cf. G. H. Hardy - J. E. Littlewood - G. Pólya 1934, Section 3.10 or J. Aczél - Z. Daróczy 1975, Section 1.3)

$$(17)\qquad \lim_{x\to +0} L(x) = L(0) = 0, \quad \lim_{x\to 1-0} L(x) = L(1) = 0 \quad \text{and } L''(x) < 0 \quad \text{on }]0,1[\ .$$

The Jensen inequality (cf. (15)) for L asserts that

$$(18)\qquad \sum_{k=1}^{N} q_k\, L(x_k) \le L(\sum_{k=1}^{N} q_k x_k)$$

holds if

$$(19)\qquad \sum_{k=1}^{N} q_k = 1;\ q_k > 0, \quad (k = 1,2,\dots,N)\ .$$

and if

$$(20)\qquad x_k \in [0,1] \quad (k = 1,2,\dots,N).$$

Again, there is equality in (18) iff

$$(6)\qquad x_1 = x_2 = \dots = x_N\ .$$

Put $x_k = p_k/q_k$ into (18) with

$$(21)\qquad p_k > 0 \quad (k = 1,2,\dots,N), \quad \sum_{k=1}^{N} p_k = 1\ .$$

By (16) we get

$$(22)\ -\sum_{k=1}^{N} q_k \frac{p_k}{q_k} \log \frac{p_k}{q_k} = \sum_{k=1}^{N} q_k\, L\,(\frac{p_k}{q_k}) \le L\,(\sum_{k=1}^{N} q_k \frac{p_k}{q_k}) = L\,(\sum_{k=1}^{N} p_k) = L\,(1) = 0$$

or

$$(8)\qquad \sum_{k=1}^{N} p_k\,(\log q_k - \log p_k) \le 0\ ,$$

equivalent to (1), with equality again iff (3) holds.

Since (18) is true on the domain (20), we may take, instead of (21),

(23) $$p_k \geq 0 \quad (k = 1,2,\dots,N), \quad \sum_{k=1}^{N} p_k = 1$$

and get from (22) at least

(24) $$\sum_{k=1}^{N} q_k \, L\left(\frac{p_k}{q_k}\right) \leq 0$$

and if, in accordance with (16) and (17), we define

(25) $$0 \log 0 := 0 ,$$

then all of (22) and so (8) and (1) hold for all p_k, q_k $(k = 1,2,\dots,N)$ satisfying (19) and (23). (Cf. for this extension also J. Aczél - J. Pfanzagl 1969.) There is still equality in (1) exactly when (3) holds. (We cannot allow

$$\sum_{k=1}^{N} p_k \leq 1$$

in (23), because L is decreasing near 1, so (24) would not hold anymore - cf. (22).)

Thus we have extended Shannon's inequality this time in another direction. We may ask whether the two extensions can be combined, that is, whether (1) is true for all p_k, q_k $(k = 1,2,\dots,N)$ satisfying

(26) $$\sum_{k=1}^{N} p_k = 1, \quad \sum_{k=1}^{N} q_k \leq 1 \; ; \; p_k \geq 0, \; q_k > 0 \quad (k = 1,2,\dots,N) .$$

The answer is yes. Indeed we have just proved (1) for p_k, q_k $(k = 1,2,\dots,N)$ satisfying (19) and (23). Suppose now that (26) is satisfied, but (19) is not, that is

(27) $$\sum_{k=1}^{N} q_k < 1 \quad (q_k > 0;\; k = 1,2,\dots,N).$$

Define

$$(28) \qquad q_{N+1} = 1 - \sum_{k=1}^{N} q_k > 0 \ , \quad p_{N+1} = 0 \ .$$

The new $p_1, p_2, \ldots, p_N, p_{N+1}, q_1, q_2, \ldots, q_N, q_{N+1}$ satisfy both (19) and (23) for N+1 instead of N, so (1) is satisfied under these circumstances, and we have (cf. (25))

$$(29) \quad -\sum_{k=1}^{N} p_k \log p_k = -\sum_{k=1}^{N} p_k \log p_k - p_{N+1} \log p_{N+1} \leq -\sum_{k=1}^{N} p_k \log q_k - p_{N+1} \log q_{N+1} =$$

$$= -\sum_{k=1}^{N} p_k \log q_k \ .$$

Thus (1) indeed holds for all p_k, q_k $(k = 1,2,\ldots,N)$ satisfying (26). Does (27) generate new equality cases? No, because

$$(30) \qquad p_{N+1} = q_{N+1}$$

would be necessary (besides $p_k = q_k$ for $k = 1,2,\ldots,N$) in order to have equality in (29), and (30) contradicts (28). We have proved the following.

<u>Theorem 1.</u> <u>The Shannon inequality</u>

$$(1) \qquad -\sum_{k=1}^{N} p_k \log p_k \leq -\sum_{k=1}^{N} p_k \log q_k$$

<u>holds (with the convention (25) where necessary), for all</u> $N \geq 2$, <u>if</u>

$$(26) \qquad \sum_{k=1}^{N} p_k = 1, \quad \sum_{k=1}^{N} q_k \leq 1; \quad p_k \geq 0, \ q_k > 0 \quad (k = 1,2,\ldots,N).$$

<u>There is equality in (1) if and only if</u>

(3) $$p_k = q_k \quad (k = 1,2,\ldots,N) .$$

As we will see, the extension (12) is the most important one for applications to optimal coding.

4. Codes are correspondances between messages and sequences of symbols, called codewords. Suppose we have D symbols and N messages $M_1, M_2, \ldots, M_N$ with the respective probabilities $p_1, p_2, \ldots, p_N$, for which (23) holds. The number of symbols in the codeword C_k corresponding to M_k $(k = 1,2,\ldots,N)$ is the length n_k of its codeword (the number of symbols in the sequence C_k). The average length of codewords in this code is

(31) $$\sum_{k=1}^{N} p_k n_k$$

The messages can be, for instance, letters. We transmit usually sequences of messages (words, in this example) by transmitting the codewords consecutively. If this can be done without spacing, that is, from the union of several codewords all of the original messages can still be uniquely determined, then we have a uniquely decipherable code. (The union of several finite sequences is the sequence obtained by writing the elements of the second sequence after those of the first, the elements of the third after those of the second sequence, and so on.) The very remarkable theorem of L. Kraft and B. McMillan (see e.g. A. Feinstein 1958, F. M. Reza 1961) announces that a code is uniquely decipherable if and only if

(32) $$\sum_{k=1}^{N} D^{-n_k} \leq 1 .$$

Often (32) is called the Kraft inequality.

It is quite natural to regard a code <u>economical</u> or <u>efficient</u>, if the average length (31) of codewords is small and <u>optimal</u> the code for which (31) is the smallest (see, however, also Section 9). But how small can (31) get? A partial answer is given in the following (see again A. Feinstein 1958 or F. M. Reza 1961 or J. Aczél - Z. Daróczy 1975, a.o.).

<u>Theorem 2.</u> <u>For every uniquely decipherable code, the average length of codewords satisfies</u>

$$(33)\qquad \sum_{k=1}^{N} p_k n_k \geq \frac{H_N(p_1,p_2,\ldots,p_N)}{\log D} \qquad \Big(\sum_{k=1}^{N} p_k = 1,\ \sum_{k=1}^{N} D^{-n_k} \leq 1;\ D \geq 2,\ n_k \text{ integers},$$

$$p_k \geq 0;\ k = 1,2,\ldots,N\Big).$$

<u>Here</u>

$$H_N(p_1,p_2,\ldots,p_N) = \sum_{k=1}^{N} L(p_k)$$

<u>with the notation (16), or, with the convention (25),</u>

$$(34)\qquad H_N(p_1,p_2,\ldots,p_N) = -\sum_{k=1}^{N} p_k \log p_k$$

<u>is the Shannon entropy of the finite probability distribution</u> $(p_1,p_2,\ldots,p_N)$, <u>if the base of the logarithms is 2.</u>

The <u>proof</u> of Theorem 2 is now very simple on basis of (32) and of Theorem 1. Indeed, put into (1)

$$(35)\qquad q_k = D^{-n_k},$$

then we get

$$H_N(p_1,p_2,\ldots,p_N) = -\sum_{k=1}^{N} p_k \log p_k \leq -\sum_{k=1}^{N} p_k \log q_k = -\sum_{k=1}^{N} p_k \log D^{-n_k} = \log D \sum_{k=1}^{N} p_k n_k$$

and this is exactly (33). The conditions (26) of Theorem 1 are satisfied, because we have supposed (23) and because (35) and the Kraft inequality (32) give

$$q_k > 0, \quad \sum_{k=1}^{N} q_k = \sum_{k=1}^{N} D^{-n_k} \leq 1 .$$

Also, by Theorem 1, we have equality in (33) iff

(36) $$p_k = q_k = D^{-n_k} \quad (k = 1,2,\ldots,N)$$

if this is possible. This is not always possible, because it means

(37) $$n_k = -\frac{\log p_k}{\log D} \quad (k = 1,2,\ldots,N)$$

and this is possible only if the fractions on the right hand sides of (37) are (positive) integers. In this case, every uniquely decipherable code with codeword lengths given by (37) is an optimal code. Also, in this sense, by (36) and (26), optimal codes are only possible for proper probability distributions with $p_k \neq 0$ $(k = 1,2,\ldots,N)$ and only if equality holds in the Kraft inequality.

If a fraction on the right hand side of (37) is not an integer, then it seems plausible that we get near to optimal codes when we choose the integers n_k near to

$$-\frac{\log p_k}{\log D} \quad (k = 1,2,\ldots,N) .$$

For instance, it is easy to prove (cf. F. M. Reza 1961, J. Aczél - Z. Daróczy 1975, also Section 9 here) that

$$(38) \qquad \frac{H_N(p_1,p_2,\ldots,p_N)}{\log D} \leq \sum_{k=1}^{N} p_k n_k < \frac{H_N(p_1,p_2,\ldots,p_N)}{\log D} + 1 ,$$

if we choose our code so (uniquely decipherable, by the Kraft - McMillan theorem) that n_k be the (unique) integer satisfying

$$(39) \qquad -\frac{\log p_k}{\log D} \leq n_k < -\frac{\log p_k}{\log D} + 1 \quad (k = 1,2,\ldots,N) .$$

There is no equality in (the first inequality of) (38) except if all

$$-\frac{\log p_k}{\log D} \quad (k = 1,2,\ldots,N)$$

are integers, i.e. if there is equality in all inequalities (39).

We can get arbitrarily small $+\varepsilon$, instead of $+1$, in (38) if we transmit sequences of independent messages consecutively. The inequalities (33) and (38) characterize the Shannon entropy (34) in a way. We will return to a generalization of this characterization in Section 9 . In the next sections we use the Shannon inequality (1), with Shannon's entropy on its left hand side, in another way for a characterization of the Shannon entropy.

5. The Shannon inequality (1) suggests the problem of determining all functions $f: \,]0,1[\, \to \mathbb{R}$ which satisfy the inequalities

$$(40) \qquad -\sum_{k=1}^{N} p_k f(p_k) < -\sum_{k=1}^{N} p_k f(q_k)$$

for all $N \geq 2$ and for all $p_1,p_2,\ldots,p_N,q_1,q_2,\ldots,q_N$ satisfying (2). The inequality (40) has also applications to the so called "how to keep the expert honest" problem, see e.g. I. J. Good 1952, J. Aczél - J. Pfanzagl 1966, J. Aczél - A. M. Ostrowski 1973. By Shannon's inequality, log is a function satisfying (40) on (2). If we multiply (40) by a nonnegative constant a and add an arbitrary constant b, we see that also the functions f given by

(41) $$f(q) = a \log q + b \quad (a \geq 0) \quad \text{for all} \quad q \in]0,1[$$

satisfy (40). (If we take the base of the logarithm arbitrary, then the constant a can be omitted, except for the trivial case a = 0.) We will prove the following.

Theorem 3. The inequality (40) or, equivalently,

(42) $$\sum_{k=1}^{N} p_k f(p_k) \geq \sum_{k=1}^{N} p_k f(q_k)$$

holds for one $N > 2$ and for all $p_1, p_2, \ldots, p_N, q_1, q_2, \ldots, q_N$ satisfying

(2) $$\sum_{k=1}^{N} p_k = \sum_{k=1}^{N} q_k = 1; \; p_k > 0, \; q_k > 0 \quad (k = 1,2,\ldots,N),$$

if and only if there exist two constants a, b such that

(41) $$f(q) = a \log q + b \quad \text{for all} \quad q \in]0,1[\quad \text{and} \quad a \geq 0$$

In this case, the left hand side of (40) is the Shannon entropy up to an additive and a nonnegative multiplicative constant.

Remarks. 1) The supposition, that (42) be satisfied for one $N > 2$, is weaker than that demanding that (42) be satisfied for all $N \geq 2$ (but also the latter is always true for (41)). If (42) is supposed only for $N = 2$ (and for all p_1, p_2, q_1, q_2 satisfying (2)), then there exist solutions different from (41), for instance $f(q) = 6q - 9q^2 + 8q^3 - 3q^4$ (for detailed discussions of the case $N = 2$, see J. Aczél - J. Pfanzagl 1966 and P. Fischer 1972).

2) Supposing (42) for all positive $q_1, q_2, \ldots, q_N$, satisfying (12) instead of (11), would again be a stronger condition for the "only if" than what we have supposed in Theorem 3. On the other hand, the "if" statement holds then too, as we have shown in Sections 2, 3. In (41), we have got the values of f on the <u>open</u> interval $]0,1[$. Indeed, we could not hope for more, since (42) under the restrictions (2) does not say anything about the values of f outside the interval $]0,1[$. If we supposed (42) valid on (23) instead of (21) (i.e. if we allowed also 0's among $p_1, p_2, \ldots, p_N$), we would get on the values of f at 1 and 0 only the restriction

$$f(1) + (N-1)\,K \geq \sup_{q \in]0,1[} f(q), \quad \text{where } 0\, f(0) := K \geq 0 .$$

3) We also do <u>not suppose</u> that there should be equality in (42) <u>only</u> in the case (3) in order to get (41). This, again, is a consequence as we have proved in Sections 2, 3.

4) Theorem 3 has been proved under the assumption of differentiability for f by J. Aczél and J. Pfanzagl 1966. P. Fischer 1972 has proved Theorem 3 in its present form, without any regularity supposition on f. However, his proof was rather difficult to understand, so several mathematicians (A. Rényi, J. Aczél, A. M. Ostrowski) have made new proofs of this remarkable theorem, see J. Aczél - A. M. Ostrowski 1973. (A similar theorem was (incorrectly) announced without proof previously by J. McCarthy 1956 with credit given to A. M. Gleason. It seems that Gleason's (unpublished) proof has been longer.) In what follows, we give two proofs of Theorem 3. While it may be difficult to recognize them, the first proof is based upon almost the same ideas as the original proof of P. Fischer 1972, although it

is, we trust, quite a bit easier to understand. The second proof is a modification and complementation of the proofs in J. Aczél - A. M. Ostrowski 1973. It utilizes also remarks made by P. Benvenuti, A. M. Gleason and W. Walter.

6. First proof of Theorem 3. We have proved the "if" part of the Theorem in Section 2. As to the "only if" part, we first show that every solution f of (42) is nondecreasing on]0,1[. We put into (42)

(43) $$p_k = q_k \text{ for all } k > 2, \quad p_1 = p, \; q_1 = q,$$

(44) $$p_1+p_2 = q_1+q_2 = r, \quad \text{i.e. } p_2 = r-p, \; q_2 = r-q$$

and (42) goes over into

(45) $$p[f(p)-f(q)] \geq (r-p)\,[f(r-q)-f(r-p)]\ .$$

Notice that r in (44) is arbitrary in]0,1[(because of (2) and (43)), so (45) holds for all

(46) $$p \in\,]0,r[,\; q \in\,]0,r[,\; r \in\,]0,1[\ .$$

The conditions (46) are symmetric in p and q, so (45) remains true on (46) if we interchange p and q:

(47) $$q[f(q)-f(p)] \geq (r-q)\,[f(r-p)-f(r-q)]\ .$$

Now multiply (45) by (r-q) and (47) by (r-p) and add these two inequalities in order to get

$$r(p-q)\,[f(p)-f(q)] = [p(r-q)-q(r-p)]\,[f(p)-f(q)] \geq 0\ .$$

This shows, that <u>for all</u> $p > q$ <u>we have</u> $f(p) \geq f(q)$, that is, f is indeed nondecreasing.

Now we show that <u>whenever</u> f <u>is differentiable at</u> r-p <u>then f is also differentiable at</u> p ($p \in]0,1[$, $r-p \in]0,1[$), <u>and</u>

$$pf'(p) = (r-p)\, f'(r-p) \tag{48}$$

<u>holds</u>. Indeed from (47) and (45), we get

$$\frac{r-q}{q} \frac{f(r-p)-f(r-q)}{(r-p)-(r-q)} \leq \frac{f(q)-f(p)}{q-p} \leq \frac{r-p}{p} \frac{f(r-p)-f(r-q)}{(r-p)-(r-q)}, \tag{49}$$

if $q > p$, and both inequalities are reversed if $q < p$. Let now $q \to p$, then $r-q \to r-p$ and both extremes of (49) tend to

$$\frac{r-p}{p} f'(r-p)$$

(since f is differentiable at (r-p)). Thus f is indeed differentiable at p and (48) holds.

What we have just proved, can also be formulated so that <u>whenever</u> f <u>is not differentiable at</u> p, <u>then</u> f <u>is also not differentiable at</u> r-p for all $r \in]p,1[$. From this it follows that f <u>is everywhere differentiable on</u> $]0,1[$. Indeed, if there existed a $p_0 \in]0,1[$ such that f were not differentiable at p_0, then f would be not differentiable at $(r-p_0)$ either, for all $r \in]p_0,1[$ that is, f would not be differentiable at any point of the interval (of positive length, $1-p_0 > 0$) $]0,1-p_0[$. But this is impossible, because f, being nondecreasing, is almost everywhere differentiable on $]0,1[$. Thus f is indeed everywhere differentiable on $]0,1[$ and <u>(48) holds for all</u> $p \in]0,1[$, $r-p \in]0,1[$, that is,

(50) $$pf'(p) = a \text{ (constant)}$$

The constant <u>a</u> in (50) is nonnegative

(51) $$a \geq 0 ,$$

because f is increasing. From (50) and (51) we get (41) (with the natural logarithm, but that makes no difference) and this concludes the first proof of Theorem 3.

<u>7</u>. <u>Second proof of Theorem 3</u>. We keep from the first proof the argument leading to the recognition that f is nondecreasing and the inequality (45), which we divide by $p-q > 0$ again

(52) $$p \, \frac{f(p)-f(q)}{p-q} \geq (r-p) \, \frac{f(r-q)-f(r-p)}{(r-q)-(r-p)} .$$

Let now q tend to p increasingly: $q \nearrow p$ (therefore $r-q \searrow r-p$). We do not know at this stage whether the two sides of (52) have limits under these circumstances, but they have (finite or infinite) lim sup's and lim inf's and the inequality in (52) remains valid between them. So we get

(53) $$p \, D^{-}f(p) \geq (r-p) \, D^{+}f(r-p)$$

and

(54) $$p \, D_{-}f(p) \geq (r-p) \, D_{+}f(r-p)$$

respectively. ($D^{-}, D^{+}, D_{-}, D_{+}$ denote the left upper, right upper, left lower, right lower Dini derivatives, respectively.) Similarly, if $p \searrow q$ ($r-p \nearrow r-q$) in (52), then we get

(55) $$q\, D^+f(q) \geq (r-q)\, D^-f(r-q) \text{ and } q\, D_+f(q) \geq (r-q)\, D_-f(r-q).$$

The inequalities (53), (54) and (55) hold, as (45), whenever

(46) $$r \in \,]0,1[\,,\ p \in \,]0,r[\,,\ q \in \,]0,r[$$

is satisfied. In particular, we may choose in (55) $q = r-p$, and then comparison with (53) and (54) gives

$$p\, D^-f(p) = (r-p)\, D^+f(r-p) \quad \text{and} \quad p\, D_-f(p) = (r-p)\, D_+f(r-p)$$

or, taking the arbitrariness of p and r within (46) into consideration, there exist two <u>(finite or infinite) constants</u> A, a such that

(56) $$x\, D^-f(x) = A = x\, D^+f(x) \quad \text{for all} \quad x \in \,]0,1[$$

and

(57) $$x\, D_-f(x) = a = x\, D_+f(x) \quad \text{for all} \quad x \in \,]0,1[\, .$$

Since f is nondecreasing, A and a must be <u>nonnegative</u>

(58) $$A \geq 0 \ , \quad a \geq 0,$$

but we have not yet ruled out the possibilities that $A = \infty$ or $a = \infty$. If we want to use, as in the first proof, the theorem that a nondecreasing function is almost everywhere differentiable or, at least, <u>differentiable at one point</u> $x_0 \in \,]0,1[$ then we get there immediately

$$D^-f(x_0) = D^+f(x_0) = D_-f(x_0) = D_+f(x_0) = f'(x_0) \ ,$$

thus A and a in (56) and (57) are equal and finite. Therefore it follows from (56) and (57) that f is differentiable at every $x \in]0,1[$:

$$\frac{a}{x} = D_-f(x) = D_+f(x) = D^-f(x) = D^+f(x) = f'(x) \quad \text{for all} \quad x \in]0,1[,$$

which is the same as (50) and since, by (58), also (51) holds, we have proved (41) and the Theorem 3 again.

In this proof, however, we will deduce Theorem 3 without appeal to the relatively deep fact that an increasing function is almost everywhere differentiable or of any other result in the theory of (Lebesgue) measure. We will also need only (57) with (58) $a \geq 0$.

First we rule out $a = \infty$. Indeed, else we would have even for arbitrarily large positive constants B

$$D_-[f(x)-Bx] = D_-f(x) - B = \infty \quad \text{and} \quad D_+[f(x)-Bx] = D_+f(x) - B = \infty$$

by (57). In particular for all functions g, defined (with different constant B's) by

$$g(x) = f(x) - Bx \qquad (x \in]0,1[),$$

we would have

(59) $\qquad D_+g(x) > 0 \;,\; D_-g(x) > 0 \quad \text{for all} \quad x \in]0,1[$.

But then g is increasing on $]0,1[$ (see Lemma 1 below). Choose, however,

$$B > 3\,[f(\tfrac{2}{3}) - f(\tfrac{1}{3})] .$$

Then

$$g(\tfrac{1}{3}) = f(\tfrac{1}{3}) - \tfrac{1}{3}B > f(\tfrac{2}{3}) - \tfrac{2}{3}B = g(\tfrac{2}{3}) ,$$

which is impossible if g is increasing. So $a = \infty$ is impossible, <u>a</u> is a finite (nonnegative) constant.

Now we can write (57) as

$$D_-[f(x)-a \ln x] = D_-f(x) - \frac{a}{x} = 0 = D_+f(x) - \frac{a}{x} = D_+[f(x)-a \ln x] \quad \text{on }]0,1[\ . \tag{60}$$

But we will prove in Lemma 2 below that a function <u>is constant on an</u> (open) <u>interval iff both its left and right lower Dini derivatives are 0 on that interval</u>. With this, Theorem 3 will be proved again, because (60) will then imply (41) (see also (58)).

<u>8</u>. We <u>prove</u> now <u>the two lemmas</u> mentioned above.

<u>Lemma 1</u>. <u>If the function</u> g <u>is defined on an (open real) interval</u> I <u>and</u>

$$D_+g(x) > 0, \ D_-g(x) > 0 \quad \text{for all} \quad x \in I, \tag{59}$$

<u>then</u> g <u>is (strictly) increasing on</u> I.

<u>Proof</u>. If, at a point $x_1 \in I$,

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

then there exists a $\delta > 0$ such that

$$\frac{g(x_1+h) - g(x_1)}{h} > \frac{k}{2} > 0 \ , \quad \text{i.e.,} \ g(x_1+h) > g(x_1) \quad \text{if} \quad 0 < h < \delta \ . \tag{61}$$

Similarly,

$$D_-g(x_1) > 0$$

implies

(62) $$g(x_1-h) < g(x_1) \quad \underline{\text{if}} \quad 0 < h < \delta \; .$$

In order to prove that g is strictly increasing we have to show, for arbitrary $x_0 \in I$, that

(63) $$g(x) > g(x_0)$$

whenever

$$x > x_0, \; x \in I \; .$$

By (59) and (61), there exist $\delta(x_0)$ such that (63) holds for all $x \in]x_0,x_0+\delta[$. Let x_1 be the **greatest** number such that (63) hold for all $x \in]x_0,x_1[\subseteq I$. This x_1 must be the right extremity of I, because else, by (61) and (62), for all sufficiently small 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 Lemma 1.

<u>Lemma 2</u>. <u>A function F is constant on an (open real) interval</u> I, <u>if and only if</u>

(64) $$D_+F(x) = D_-F(x) = 0 \quad \underline{\text{for all}} \quad x \in I \; .$$

<u>Proof</u>. The "only if" part is obvious. In order to prove the "if" part, define first g by

(65) $$g(x) = F(x) + \varepsilon x \quad (\varepsilon > 0) \quad \text{for all} \quad x \in I \; .$$

For this function, by (64) and (65),

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

Thus, by Lemma 1, g is increasing on I. Therefore ($\varepsilon \to 0$, cf. (65)), F is nondecreasing on I.- We prove now that F, being already monotonic, is also continuous on I. Indeed, all discontinuities of monotonic functions are jumps, and there either D_+F or D_-F would be ∞ , contrary to (64).

Thus F is continuous and nondecreasing. We conclude the proof of Lemma 2 by showing that there do not exist

$$\text{(66)} \qquad a \in I,\ b \in I,\ a < b \quad \text{such that} \quad F(a) < F(b) ,$$

thus showing that F is constant on I. Indeed, if (66) held, we would define the linear function ℓ by

$$\text{(67)} \qquad \ell(x) = F(a) + \varepsilon(x-a) \quad (\varepsilon > 0) \quad \text{for all} \quad x \in I .$$

We have $\ell(a) = F(a)$ but, if ε is small enough, then $\ell(b) = F(a) + \varepsilon\ (b-a) < F(b)$, by (66). The function F being continuous, there would exist a greatest $x_1 \in [a,b[$ for which $F(x_1) = \ell(x_1)$, while, for all $x \in]x_1,b]$, $F(x) > \ell(x)$. But then

$$\text{(68)} \qquad \frac{F(x) - F(x_1)}{x - x_1} > \frac{\ell(x) - \ell(x_1)}{x - x_1} \ (x > x_1) \quad \text{and} \quad D_+F(x_1) \geq \ell'(x_1) = \varepsilon > 0$$

would hold, contrary to (64). Thus Lemma 2 is proved.

Remark 5) One has to be careful with plausible sounding things about Dini derivatives. For instance, the statement in Lemma 1 (that g is increasing on I) is not true, if only

(69) $$D_+g(x) > 0 \quad \text{on } I$$

is supposed. Counter example: $I =]0,1[$,

$$g(x) = \begin{cases} x, & \text{if } x \in]0,\frac{1}{2}[\\ x-1, & \text{if } x \in [\frac{1}{2},1[\ . \end{cases}$$

The function g is increasing, however, if, besides (69), also the continuity of g is supposed on I. The situation is analogous for Lemma 2.

Also, in general it is not true that

$$D_+(F+G) = D_+F+D_+G \ ,$$

only

$$D_+(F+G) \geq D_+F+D_+G \ ,$$

but, first, this would be enough for the above proofs and, second,

$$D_+(F+G) = D_+F+G'$$

if G is differentiable.

9. We return now to the relationships among optimal coding, entropies and Shannon-type inequalities. We have pointed out in Section 4, that Shannon's inequality establishes a connection between Shannon's entropy

and optimal coding, if the optimality of coding is measured by minimizing the arithmetic average length (31) of a codeword. We call this average length arithmetic, because (31) is the arithmetic mean of the lengths $n_1, n_2, \ldots, n_N$ of codewords, weighted with the probabilities $p_1, p_2, \ldots, p_N$. There exist also other mean values (e.g. the geometric mean in (4)).

L. L. Campbell (1965) has suggested the use of the

$$(70) \qquad \frac{1}{t} \log_D \left(\sum_{k=1}^{N} p_k D^{tn_k} \right) \quad (t \neq 0 \; ; \; \sum_{k=1}^{N} p_k = 1; \; p_k > 0; \; k = 1,2,\ldots,N)$$

exponential mean lengths of codewords, weighted again with the probabilities ($\log_D$ is the logarithm with base D, of course). In order to get a result similar to Theorem 2, we use another inequality instead of the (1) Shannon inequality.

Hölder's inequality (G. H. Hardy - J. E. Littlewood - G. Pólya 1934, Section 2.8) states that

$$(71) \qquad \sum_{k=1}^{N} x_k y_k \geq \left(\sum_{k=1}^{N} x_k^p \right)^{1/p} \left(\sum_{k=1}^{N} y_k^q \right)^{1/q}, \; \underline{\text{if}} \; \frac{1}{p} + \frac{1}{q} = 1 \quad \underline{\text{and if}} \quad p < 1, \; p \neq 0,$$

while

$$(72) \qquad \sum_{k=1}^{N} x_k y_k \leq \left(\sum_{k=1}^{N} x_k^p \right)^{1/p} \left(\sum_{k=1}^{N} y_k^q \right)^{1/q}, \; \underline{\text{if}} \; \frac{1}{p} + \frac{1}{q} = 1 \quad \underline{\text{and if}} \quad p > 1 \, .$$

Here $x_k > 0$, $y_k > 0$; $k = 1,2,\ldots,N$ (later we will allow zeros among the x's and y's). There is equality in (71) and (72) iff the sequences $\{x_k\}$, $\{y_k\}$ are proportional, that is, there exists a (positive) number c such that

$$(73) \qquad x_k^p = c \, y_k^q \, .$$

In (71), put

$$x_k = p_k^{-1/t} D^{-n_k}, \quad y_k = p_k^{1/t}, \; p = -t \quad (\text{and so } q = \frac{p}{p-1} = \frac{t}{t+1}) . \tag{74}$$

Because of $p < 1$, $p \neq 0$, we have

$$t > -1 \; , \quad t \neq 0 \; . \tag{75}$$

Thus we get from (71), taking also the (32) Kraft inequality into consideration,

$$1 \geq \sum_{k=1}^{N} D^{-n_k} = \sum_{k=1}^{N} x_k y_k \geq \left(\sum_{k=1}^{N} p_k D^{tn_k}\right)^{-1/t} \left(\sum_{k=1}^{N} p_k^{1/(t+1)}\right)^{(t+1)/t} \tag{76}$$

or

$$\left(\sum_{k=1}^{N} p_k D^{tn_k}\right)^{1/t} \geq \left(\sum_{k=1}^{N} p_k^{1/(t+1)}\right)^{(t+1)/t} ,$$

$$\frac{1}{t} \log_D \left(\sum_{k=1}^{N} p_k D^{tn_k}\right) \geq \frac{t+1}{t} \log_D \left(\sum_{k=1}^{N} p_k^{1/(t+1)}\right) . \tag{77}$$

We have now the expression (70) on the left hand side of (77). As to the right hand side, if we introduce

$$\alpha := \frac{1}{t+1} , \tag{78}$$

where

$$\alpha > 0, \; \alpha \neq 1 \tag{79}$$

by (75), then (77) goes over into

(80) $$\frac{\alpha}{1-\alpha}\log_D\left(\sum_{k=1}^{N} p_k D^{n_k(1-\alpha)/\alpha}\right) \geq \frac{1}{1-\alpha}\log\left(\sum_{k=1}^{N} p_k^{\alpha}\right)/\log D .$$

This now is an inequality similar to (33). Indeed, the left hand side is an (exponential) mean length of codewords (cf. (70), (78)) while the quantities

(81) $${}_{\alpha}H_N(p_1,p_2,\ldots,p_N) = \frac{1}{1-\alpha}\log\left(\sum_{k=1}^{N} p_k^{\alpha}\right) \quad (\alpha \neq 1;\ \sum_{k=1}^{N} p_k = 1;\ p_k \geq 0;\ k = 1,2,\ldots,N)$$

(log again with base 2) on the right hand side of (80) are the <u>entropies of order</u> α introduced by A. Rényi in 1960 and called <u>Rényi entropies</u>. We call the left hand side of (80)

(82) $$\frac{\alpha}{1-\alpha}\log_D\left(\sum_{k=1}^{N} D^{n_k(1-\alpha)/\alpha}\right), \quad (\alpha \neq 1,\ \alpha \neq 0)$$

an α-<u>average length of codewords</u>. With (81), we can write (80) as

(83) $$\frac{\alpha}{1-\alpha}\log_D\left(\sum_{k=1}^{N} p_k D^{n_k(1-\alpha)/\alpha}\right) \geq \frac{{}_{\alpha}H_N(p_1,p_2,\ldots,p_N)}{\log D}, \quad (\alpha > 0,\ \alpha \neq 1;$$

$$\sum_{k=1}^{N} p_k = 1;\ p_k > 0;\ k = 1,2,\ldots,N),$$

where the similarity to (33) is even easier to recognize. As a matter of fact, the following is also easy to prove. If $\alpha \to 1$, then (82) ((70) if $t \to 0$) tends to (31), while the (81) Rényi entropy of order α tends to the Shannon entropy (34). So, we have <u>just given a new proof of Theorem 2</u>. We had till now $p_k > 0$; $k = 1,2,\ldots,N$; but, by (79) and (80) the same remains true also if some p_k's are zero.

We may thus call <u>the Shannon entropy a Rényi entropy of order</u> 1, and (31) a 1-<u>average length of codewords</u>. Summarizing we have proved the following.

Theorem 4. For every uniquely decipherable code, the α-average length of codewords satisfies

(84) $$\frac{\alpha}{1-\alpha}\log_D\left(\sum_{k=1}^{N} p_k D^{n_k(1-\alpha)/\alpha}\right) \geq \frac{{}_\alpha H_N(p_1,p_2,\ldots,p_N)}{\log_2 D} \quad \text{whenever} \quad \alpha > 0 ;$$

$$D \geq 2,\ n_k \text{ integers},\ p_k \geq 0\ (k = 1,2,\ldots,N);\ \sum_{k=1}^{N} D^{-n_k} \leq 1,\ \sum_{k=1}^{N} p_k = 1 ,$$

where ${}_\alpha H_N(p_1,p_2,\ldots,p_N)$ is the (81) Rényi entropy if $\alpha \neq 1$ and the (34) Shannon entropy if $\alpha = 1$, while the left hand side of (84) is replaced by

(31) $$\sum_{k=1}^{N} p_k n_k ,$$

if $\alpha = 1$.

Let us see, when is equality possible in (83). Only if there is equality in both inequalities of (76). The first equality means

(85) $$\sum_{k=1}^{N} D^{-n_k} = 1 ,$$

the second can, by (73), (74), and (78), hold iff

$$D^{n_k(1-\alpha)/\alpha} = c\, p_k^{\alpha-1}\ (k = 1,2,\ldots,N) ,$$

i.e.,

(86) $$D^{-n_k} = c^{\alpha/(\alpha-1)} p_k^{\alpha}\ (k = 1,2,\ldots,N) .$$

Combined with (85), we would have

$$c^{\alpha/(\alpha-1)} = 1/\sum_{k=1}^{N} p_k^{\alpha}$$

and, from (86)

(87) $$n_k = -\log_D\left(p_k^{\alpha}/\sum_{j=1}^{N} p_j^{\alpha}\right)\ (k = 1,2,\ldots,N;\ \alpha > 0) .$$

This is possible only if the quantities on the right hand side of (87) are (positive) integers. In this case, every uniquely decipherable code with codeword lengths given by (87) is an optimal code in the sense that it minimizes the α-average lengths of codewords. In particular, as we have seen, such optimal codes are only possible if equality holds in the Kraft inequality.

Here too, if some of the right sides of (87) are not integers, then it seems plausible that we get near to optimal codes when we choose the integers n_k near to

$$-\log_D \left(p_k^{\alpha} / \sum_{j=1}^{N} p_j^{\alpha}\right) \quad (k = 1,2,\ldots,N) .$$

In particular, we have the following (cf. L. L. Campbell 1965, J. Aczél - Z. Daróczy 1975).

Theorem 5. There exist uniquely decipherable codes for which the inequalities

$$(88)\qquad \frac{{}_{\alpha}H_N(p_1,p_2,\ldots,p_N)}{\log_2 D} \le \frac{\alpha}{1-\alpha} \log_D \left(\sum_{k=1}^{N} p_k D^{n_k(1-\alpha)/\alpha}\right) < \frac{{}_{\alpha}H_N(p_1,p_2,\ldots,p_N)}{\log D} + 1 \quad (\alpha >$$

hold (in the case $\alpha = 1$, $\sum_{k=1}^{N} p_k n_k$ stands in the middle). We get such codes if we choose the codeword lengths n_k as the (unique) integers satisfying

$$(89)\qquad -\log_D \left(p_k^{\alpha} / \sum_{j=1}^{N} p_j^{\alpha}\right) \le n_k < -\log_D \left(p_k^{\alpha} / \sum_{j=1}^{N} p_j^{\alpha}\right) + 1 \quad (k = 1,2,\ldots,N) .$$

There is equality in (the first inequality of) (88) exactly when all quantities

$$-\log_D \left(p_k^{\alpha} / \sum_{j=1}^{N} p_j^{\alpha}\right) \quad (k = 1,2,\ldots,N)$$

are integers, that is when there is equality in all (left) inequalities (89).

This time we prove (88), which is analogous to, and contains as (limiting) case $\alpha = 1$, the inequalities (38).

From the left inequalities of (89)

$$D^{-n_k} \leq \frac{p_k^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \qquad (k = 1,2,\ldots,N)$$

follows. If we take the sums of both sides from $k = 1$ to $k = N$, we get that the

$$(32) \qquad \sum_{k=1}^{N} D^{-n_k} \leq 1$$

Kraft inequality is satisfied, that is, there indeed exist uniquely decipherable codes with the codeword lengths determined by (89). We have seen in Theorems 2 and 4 that these satisfy the first inequality of (88). In order to prove that they satisfy also the second inequality (88), we introduce the notation

$$(90) \qquad {}_\alpha M(\{x_k\}) = \frac{\alpha}{1-\alpha} \log_D \left(\sum_{k=1}^{N} p_k D^{x_k(1-\alpha)/\alpha} \right) \quad (\alpha(\alpha-1) \neq 0)$$

($\alpha > 0$; $x_1, x_2, \ldots, x_N$ is arbitrary real but

$$\sum_{k=1}^{N} D^{-x_k} \leq 1 \, ;$$

in particular, (82) is ${}_\alpha M(\{n_k\})$). We note that the exponential mean (90) ${}_\alpha M$ is increasing with each x_k $(k = 1,2,\ldots,N)$ and translatory, i.e.

(91) $${}_{\alpha}M(\{x_k+t\}) = {}_{\alpha}M(\{x_k\}) + t \quad \text{for all real} \quad x_1, x_2, \ldots, x_N, \ t \ .$$

Then the right inequalities in (89) imply

(92) $$\frac{\alpha}{1-\alpha} \log \left(\sum_{k=1}^{N} p_k \, D^{n_k(1-\alpha)/\alpha} \right) = {}_{\alpha}M(\{n_k\}) < {}_{\alpha}M(\{-\log_D(p_k^{\alpha}/\sum_{j=1}^{N} p_j^{\alpha}) + 1\}) =$$

$$= {}_{\alpha}M(\{-\log_D(p_k^{\alpha}/\sum_{j=1}^{N} p_j^{\alpha})\}) + 1 = \frac{\alpha}{1-\alpha} \log_D \left(\sum_{k=1}^{N} p_k D^{-\log_D(p_k^{\alpha}/\sum_{j=1}^{N} p_j^{\alpha})\,(1-\alpha)/\alpha} \right) + 1 =$$

$$= \frac{\alpha}{1-\alpha} \log_D \left[\sum_{k=1}^{N} p_k p_k^{\alpha(\alpha-1)/\alpha} \left(\sum_{j=1}^{N} p_j^{\alpha} \right)^{(1-\alpha)/\alpha} \right] + 1 =$$

$$= \frac{\alpha}{1-\alpha} \log_D \left[\left(\sum_{k=1}^{N} p_k^{\alpha} \right) \left(\sum_{k=1}^{N} p_k^{\alpha} \right)^{(1/\alpha)-1} \right] + 1 =$$

$$= 1 + \frac{1}{1-\alpha} \log \left(\sum_{k=1}^{N} p_k^{\alpha} \right) / \log D = \frac{{}_{\alpha}H_N(p_1, p_2, \ldots, p_N)}{\log D} + 1 \ ,$$

as asserted. It seems that we have excluded in (92) case $\alpha = 1$ (Shannon entropies, average lengths (31) of codewords). But, evidently, also the arithmetic mean ${}_{1}M$ defined by

(93) $${}_{1}M(\{x_k\}) = \sum_{k=1}^{N} p_k x_k$$

is increasing and satisfies the translativity (91). The quantity (31) is again ${}_{1}M(\{n_k\})$. Now (92) remains valid, with the appropriate changes for $\alpha = 1$. The rest of Theorem 5 is obvious.

We can again get arbitrarily small $+ \varepsilon$, instead of $+ 1$, in (88) if we transmit sequences of independent messages.

10. The inequalities (84) and (88) characterize the Shannon and Rényi entropies of order $\alpha(>0)$. But what characterizes the α-average lengths (31) and (82) which figure in these inequalities or the arithmetic and exponential means (93) and (90)? We try to answer this question in this last Section.

We recall the proof of Theorem 5. Both (90) and (93) are quasiarithmetic means, which means that they are of the form

$$(94)\qquad {}^{\psi}M(\{x_k\}, \{p_k\}) = \psi^{-1}\left(\sum_{k=1}^{N} p_k\,\psi(x_k)\right) \quad \left(\sum_{k=1}^{N} p_k = 1;\ p_k > 0;\ k = 1,2,\ldots,N;\ \sum_{k=1}^{N} D^{-x_k} \leq 1\right)$$

where ψ is a continuous and strictly monotonic function. For (93) and (90),

$$(95)\qquad \psi(x) = x \quad \text{or} \quad \psi(x) = D^{x(1-\alpha)/\alpha} \qquad (\alpha \neq 0,\ \alpha \neq 1),$$

respectively. Mean values of the form (94) are evidently always increasing in the x_k's $(k = 1,2,\ldots,N)$. But when are they also

$$(96)\qquad {}^{\psi}M(\{x_k+t\}, \{p_k\}) = {}^{\psi}M(\{x_k\}, \{p_k\}) + t$$

translatory (cf. (91), a property which we also needed in (92)). It is known (see e.g. J. Aczél 1966, cf. also G. H. Hardy - J. E. Littlewood - G. Pólya 1934, Section 3.3) that the only translatory (96) quasiarithmetic means are the arithmetic and exponential means (93) and (90) with $\alpha \neq 0$.

We have still to motivate (96) from the information theoretic point of view and also exclude $\alpha < 0$ in (90).

Remember, that we have introduced (93) and (90) as generalizations of the α-average lengths (31) and (82) of codewords. Take two independent sets of messages $K_1,K_2,\ldots,K_L$ and $M_1,M_2,\ldots,M_N$ with the respective probabilities

$p_1, p_2, \ldots, p_L, q_1, q_2, \ldots, q_N$ and codeword lengths $\ell_1, \ell_2, \ldots, \ell_L, n_1, n_2, \ldots, n_N$ in their respective uniquely decipherable codes (same number D of symbols in both). Then there exist uniquely decipherable codes with codeword lengths $\ell_k + n_m$ for coding pairs of messages (K_k, M_m) $(k = 1,2,\ldots,L;\ m = 1,2,\ldots,N)$. Indeed from the Kraft inequalities (cf. (32))

$$\sum_{k=1}^{L} D^{-\ell_k} \leq 1 \quad \text{and} \quad \sum_{m=1}^{N} D^{-n_m} \leq 1,$$

also the Kraft inequality

$$\sum_{(k,m)=(1,1)}^{(L,N)} D^{-\ell_k + n_m} \leq 1$$

follows, by multiplication. Since the messages K_k, M_m $(k = 1,2,\ldots,L;$ $m = 1,2,\ldots,N)$ were supposed independent, the probability of the pair (K_k, M_m) will be $p_k q_m$ $(k = 1,2,\ldots,L;\ m = 1,2,\ldots,N)$. In analogy to (31), (82) and (94), we can introduce quasiarithmetic mean lengths of codewords by

$$(97) \qquad {}^{\psi}M(\{\ell_k\}, \{p_k\}) = \psi^{-1}\left[\sum_{k=1}^{L} p_k\, \psi(\ell_k)\right],$$

and it is quite natural to ask which of these are additive, i.e.,

$$(98) \qquad {}^{\psi}M(\{\ell_k + n_m\}, \{p_k q_m\}) = {}^{\psi}M(\{\ell_k\}, \{p_k\}) + {}^{\psi}M(\{n_m\}, \{q_m\}) .$$

This means that the mean length of codewords in the code for pairs of messages should be equal to the sum of mean lengths of codeword in the codes for individual messages.

It is easy to check that the arithmetic and exponential mean codeword lengths (31) and (82) have this additive property (98), but the problem of determining all quasiarithmetic mean lengths (97) of codewords, which are (98) additive, is still unsolved.(*)

However, L. L. Campbell (1966, see also J. Aczél - Z. Daróczy 1975) has introduced noninteger codeword lengths and stated that they can be motivated also from the point of view of coding theory. (One advantage is that the lower bounds in (33) and (84) can then be actually attained.) Then (97) is defined (cf. (94)) and (98) postulated for all real (or all positive) ℓ_k, n_m with (32), and $p_k > 0$, $q_m > 0$ $(k = 1,2,\ldots,L;\ m = 1,2,\ldots,N)$ with $\sum_{k=1}^{L} p_k = \sum_{m=1}^{N} q_m = 1$. For these, the above question is now easily solved since, under these circumstances, (96) follows from (98) if we take $n_m = t$ $(m = 1,2,\ldots,N)$, and we have just seen that the expressions (93) and (90) (with $\alpha \neq 0$) are the only quasiarithmetic means which satisfy (96). It still remains to motivate the restriction $\alpha > 0$.

Remember, that the inequality (84) was our primary reason for introducing α-average codeword lengths. Both sides (cf. also (81)) make sense also if $\alpha < 0$. But the inequality does not hold in general, if $\alpha < 0$. Take, for instance, $\alpha = -1$, $p_1 = \frac{1}{3}$, $p_2 = \frac{2}{3}$, $D = 2$, $n_1 = n_2 = 1$ (the (32) Kraft inequality is satisfied). Then

$$\frac{{}_{-1}H_2(\frac{1}{3},\frac{2}{3})}{\log_2 2} = \frac{1}{2}\log_2\left[\left(\frac{1}{3}\right)^{-1} + \left(\frac{2}{3}\right)^{-1}\right] = \frac{1}{2}\log_2 4.5 > \frac{1}{2}\log_2 4 = 1 =$$

$$= -\frac{1}{2}\log_2\left(\frac{1}{3}2^{-2} + \frac{2}{3}2^{-2}\right) = {}_{-1}M(\{1\}),$$

i.e., there is $<$ instead of $\geq$ in (84).

As a matter of fact, we have seen in (76) that the proof of (84) is based upon the (32) Kraft inequality and the (71) Hölder inequality. If there is equality in (32) and if the Hölder inequality is reversed with strict $<$ instead of $\geq$, then we have in (84) $<$ instead of $\geq$. The former condition is satisfied, for instance, if

(*) Since completion of this paper, I have solved this problem. The solutions are (31) and (82).

(99) $$N = D^n,\ n_k = n \quad (k = 1,2,\ldots,D).$$

The latter condition is satisfied, see (72), if $p > 1$ and $\{x_k^p\}$ is not proportional to $\{y_k^q\}$. That means, see (74), (78) and (99),

$$p_k D^{-np} = x_k^p \neq c\, y_k^q = c\, p_k^{1/(1-p)} \quad \text{i.e. } p_k \neq c^{(p-1)/p} D^{n(p-1)} \quad (k = 1,2,\ldots,N) \text{ and } t < -1,$$

that is

(100) $$\alpha < 0 \quad \text{and } (p_1,p_2,\ldots,p_N) \neq (\tfrac{1}{N},\tfrac{1}{N},\ldots,\tfrac{1}{N})$$

So we have proved that <u>in (84)</u> < <u>stands instead of</u> ≥ <u>if (99) and (100) hold</u>.

With aid of the functions given in (95), the Shannon and Rényi entropies (34) and (81), when divided by $\log_2 D$ as in (33) and (83), can be written as

(101) $$^{\psi}H_N^D(\{p_k\}) = (\log_D \psi(1) + 1)\ \psi^{-1}\left[\sum_{k=1}^{N} p_k\ \psi(-\log_D p_k)^{1/(\log_D \psi(1)+1)}\right]$$

Summarizing, we have proved the following.

<u>Theorem 6</u>. <u>The</u> α-<u>average generalized codeword lengths (90) and (93)</u>

(102) $$\frac{\alpha}{1-\alpha} \log_D\left(\sum_{k=1}^{N} p_k D^{x_k(\alpha-1)/\alpha}\right) \quad (\alpha \neq 0,\ \alpha \neq 1),\quad \sum_{k=1}^{N} p_k x_k \quad (\alpha = 1)$$

<u>and only these are quasiarithmetic</u>

$$^{\psi}M_N(\{x_k\},\ \{p_k\}) = \psi^{-1}\left(\sum_{k=1}^{N} p_k \psi(x_k)\right) \quad \left(\sum_{k=1}^{N} p_k = 1,\ \sum_{k=1}^{N} D^{-x_k} \leq 1;\right.$$

$p_k > 0$, x_k <u>real</u>; $k = 1,2,\ldots,N$)

(ψ continuous and strictly monotonic) and (98) additive

$$^{\psi}M_{LN}(\{x_k+y_m\}, \{p_k q_m\}) = {}^{\psi}M_L(\{x_k\}, \{p_k\}) + {}^{\psi}M_N(\{y_m\}, \{q_m\}) .$$

If these quasiarithmetic mean generalized codeword lengths are also bounded from below by the respective entropies (101)

$$^{\psi}M_N(\{x_k\}, \{p_k\}) \geq {}^{\psi}H_N^N(\{p_k\})$$

for at least one $N \geq 2$, one (finite) probability distribution with

$$(p_1, p_2, \ldots, p_N) \neq (\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N})$$

and with $n_k = 1$ $(k = 1,2,\ldots,N)$, then $\alpha > 0$ in (102).

References

J. Aczél 1966, Lectures on Functional Equations and Their Applications. Academic Press, New York - London, Section 3.1.

J. Aczél - Z. Daróczy 1975. On Information Measures and Their Characterizations. Academic Press, New York (in press).

J. Aczél - A. M. Ostrowski 1973, On the Characterization of Shannon's Entropy by Shannon's Inequality. J. Austral. Math. Soc. 16, 368-374.

J. Aczél - J. Pfanzagl 1966, Remarks on the Measurement of Subjective Probability and Information. Metrika 11, 91-105.

L. L. Campbell 1965, A Coding Theorem and Rényi's Entropy, Information and Control 8, 423-429.

L. L. Campbell 1966, Definition of Entropy by Means of a Coding Problem. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 6, 113-118.

A. Feinstein 1958, Foundations of Information Theory. McGraw-Hill, New York - Toronto - London, Chapter 2.

P. Fischer 1972, On the Inequality $\Sigma p_i f(p_i) \geq \Sigma p_i f(q_i)$. Metrika 18, 199-208.

I. J. Good 1952, Rational Decisions. J. Roy. Statist. Soc. Ser. B 14, 107-114.

G. H. Hardy - J. E. Littlewood - G. Pólya 1934, Inequalities. University Press, Cambridge (2nd ed. 1952).

J. McCarthy 1956, Measures of the Value of Information. Proc. Nat. Acad. Sci. U.S.A. 42, 654-655.

A. Rényi 1960, On Measures of Entropy and Information. Proc. 4th Berkeley Sympos. Math. Statist. and Prob. 1960, Vol. I. University of California Press, Berkeley, Calif. 1961, pp.547-561.

F. M. Reza 1961, An Introduction to Information Theory. McGraw-Hill, New York - Toronto - London, Chapter 4.

2. Determination of All Additive Quasiarithmetic Mean Codeword Lengths

J. Aczél

Determination of All Additive Quasiarithmetic Mean Codeword Lengths

J. Aczél

1. L. L. Campbell 1966 has introduced quasiarithmetic mean codeword lengths in the following manner.

Let $Y = \{\eta_1, \eta_2, \ldots, \eta_K\}$ be a finite set of messages and let $Q = \{q_1, q_2, \ldots, q_K\}$ be an associated distribution of probabilities, so that the probability of η_k is q_k $(k = 1,2,\ldots,K)$ and

$$(1) \qquad \sum_{k=1}^{K} q_k = 1; \; q_k \geq 0 \quad (k = 1,2,\ldots,K) .$$

Suppose that we wish to represent the messages in Y by codewords, i.e. by finite sequences of elements of the set $\{0,1,\ldots,D-1\}$ where $D > 1$. There is a uniquely decipherable code (see e.g. F. M. Reza 1961) which represents η_k by a codeword of length (number of elements) n_k $(k = 1,2,\ldots,K)$ if and only if the set of positive integer codeword lengths $N = \{n_1, n_2, \ldots, n_K\}$ satisfies the Kraft inequality

$$(2) \qquad \sum_{k=1}^{K} D^{-n_k} \leq 1 .$$

Let now ϕ: $[1,\infty[\to R$ be a continuous strictly increasing function. It has an inverse ϕ^{-1} which is also continuous and strictly increasing. This defines a quasiarithmetic mean codeword length

$$(3) \qquad L(Q,N;\phi) = \phi^{-1}\Big[\sum_{k=1}^{K} q_k \phi(n_k)\Big]$$

for all N satisfying (2). The reason for calling L a mean length is that, for $N = \{n,n,\dots,n\}$, i.e. when all codewords are of equal length n, then $L(Q,N;\phi) = n$. Moreover, if $\phi(x) = \phi_0(x) = x$ $(x \in [1,\infty[)$, then

$$L(Q,N;\phi) = \sum_{k=1}^{K} q_k n_k , \tag{4}$$

the ordinary or arithmetic mean codeword length. L. L. Campbell 1965, 1966 has also introduced the exponential mean codeword length, for which $\phi(x) = \phi_t(x) = D^{tx}$ $(x \in [1,\infty[;\ t \neq 0)$,

$$L(Q,N;\phi_t) = \frac{1}{t} \log_D \sum_{k=1}^{K} q_k D^{tn_k} . \tag{5}$$

It is easy to see that $\lim_{t\to 0} L(Q,N;\phi_t) = L(Q,N;\phi_0)$.

Important inequalities are known for the mean codeword lengths (4) and (5) (see, e.g., F. M. Reza 1961, L. L. Campbell 1965, J. Aczél 1973 and section 4 of the present paper). These give essentially the Shannon and Rényi entropies as lower bounds of (4) and (5), respectively, and show also that there exist uniquely decipherable codes for which these mean codeword lengths come within a unit (bit) from their lower bounds. The proof of the latter facts use a translativity property of (4) and (5), the generalizations of which we will examine in section 3. The inequalities, mentioned above, can also be translated into optimal coding statements with respect to certain cost functions, related to ϕ in (3). This we will see in section 4, in modification of results by L. L. Campbell, partly published (L. L. Campbell 1965, 1966) and partly unpublished.

The question arises, why the mean codeword lengths (4) and (5) have been chosen, say, among the quasiarithmetic mean codeword lengths (3). In our main result, in section 2, we will show that the following rather natural additivity condition characterizes them.

Consider two independent sets of messages $X = \{\xi_1,\xi_2,\ldots,\xi_J\}$ and $Y = \{\eta_1,\eta_2,\ldots,\eta_K\}$ with associated probability distributions $P = \{p_1,p_2,\ldots,p_J\}$ and $Q = \{q_1,q_2,\ldots,q_K\}$. Since X and Y are independent, the probability of the pair (ξ_j,η_k) is p_jq_k $(j = 1,2,\ldots,J;$ $k = 1,2,\ldots,K)$. We denote by PQ the probability distribution $\{p_1q_1,p_1q_2,\ldots,p_1q_K,p_2q_1,p_2q_2,\ldots,p_2q_K,\ldots,p_Jq_1,p_Jq_2,\ldots,p_Jq_K\}$. Let ξ_j be represented by a codeword of length m_j $(j = 1,2,\ldots,J)$ and let η_k be represented by a codeword of length n_k $(k = 1,2,\ldots,K)$. Moreover, suppose that we use the same symbols $\{0,1,\ldots,D-1\}$ in all these representations. The pair (ξ_j,η_k) may be represented by a codeword of length $m_j + n_k$ $(j = 1,2,\ldots,J;\ k = 1,2,\ldots,K)$. Let us denote these three distributions of lengths by $M = \{m_1,m_2,\ldots,m_J\}$, $N = \{n_1,n_2,\ldots,n_K\}$ and

$$M + N = \{m_1+n_1,m_1+n_2,\ldots,m_1+n_K,m_2+n_1,m_2+n_2,\ldots,m_2+n_K,\ldots,m_J+n_1,m_J+n_2,\ldots,m_J+n_K\},$$

respectively. If M and N satisfy the Kraft inequality (2) then so does M + N because

$$\sum_{j=1}^{J} D^{-m_j} \le 1 \quad \text{and} \quad \sum_{k=1}^{K} D^{-n_k} \le 1 \tag{6}$$

imply

$$\sum_{j=1}^{J}\sum_{k=1}^{K} D^{-(m_j+n_k)} \le 1 .$$

Thus there exists indeed a uniquely decipherable code with M + N as set of codeword lengths for $XY = \{\xi_1\eta_1,\xi_1\eta_2,\ldots,\xi_1\eta_K,\xi_2\eta_1,\xi_2\eta_2,\ldots,\xi_2\eta_K,\ldots,\xi_J\eta_1,\xi_J\eta_2,\ldots,\xi_J\eta_K\}$.

If L is to be a measure of mean lengths, it is natural to require that

(7) $$L(PQ,M+N;\phi) = L(P,M;\phi) + L(Q,N;\phi) ,$$

i.e.,

(8) $$\phi^{-1}\Big[\sum_{j=1}^{J}\sum_{k=1}^{K} p_j q_k \phi(m_j+n_k)\Big] = \phi^{-1}\Big[\sum_{j=1}^{J} p_j \phi(m_j)\Big] + \phi^{-1}\Big[\sum_{k=1}^{K} q_k \phi(n_k)\Big].$$

We call the properties (7) or (8) additivity. They are supposed for all positive integers m_j and n_k satisfying (6) and for all p_j, q_k $(j = 1,2,\ldots,J;\ k = 1,2,\ldots,K)$ satisfying (1) and

(9) $$\sum_{j=1}^{J} p_j = 1;\ p_j \geq 0 \quad (j = 1,2,\ldots,J) .$$

The problem of finding all additive (7), quasiarithmetic (3) mean codeword lengths has not been solved before (cf. L. L. Campbell 1966, J. Aczél 1973). Instead, L. L. Campbell 1966, has generalized the codeword lengths n_k $(k = 1,2,\ldots,K)$ so that they become arbitrary real numbers satisfying (2), and has solved (8) in this case. In this paper we solve the original problem, with positive integer codeword lengths. We restrict ourselves to $J = K = 2$, thus making the result more general. This has also the advantage that, because of $D \geq 2$, $m_1 \geq 1$, $m_2 \geq 1$, $n_1 \geq 1$, $n_2 \geq 1$, (6) is always satisfied.

2. Theorem 1. **The arithmetic and the exponential mean codeword lengths (4) and (5) are the only quasiarithmetic mean codeword lengths (3) which are additive (7) with** $J = K = 2$ **(for two-place distributions).**

Proof. For $J = K = 2$, (7) or (8) can be written as

(10) $$\phi^{-1}[p_1q_1\phi(m_1+n_1) + p_1q_2\phi(m_1+n_2) + p_2q_1\phi(m_2+n_1) + p_2q_2\phi(m_2+n_2)] =$$

$$= \phi^{-1}[p_1\phi(m_1) + p_2\phi(m_2)] + \phi^{-1}[q_1\phi(n_1) + q_2\phi(n_2)]$$

where

(11) $$p_1 \geq 0,\ p_2 \geq 0,\ p_1+p_2 = 1,\ q_1 \geq 0,\ q_2 \geq 0,\ q_1+q_2 = 1,$$
$$m_1, m_2, n_1, n_2 \text{ are positive integers.}$$

Put into (10) $m_1 = m_2 = m$, $q_1 = 1-q$, $q_2 = q$, in order to get

(12) $$\phi^{-1}[(1-q)\phi(n_1+m) + q\phi(n_2+m)] = \phi^{-1}[(1-q)\phi(n_1) + q\phi(n_2)] + m$$

for all

(13) $$q \in [0,1];\quad n_1,\ n_2,\ m \text{ positive integers.}$$

We need the following

Lemma. Let ϕ, ψ be continuous, strictly increasing functions defined on $[1,\infty[$. The equation

(14) $$\phi^{-1}[(1-q)\phi(n_1) + q\phi(n_2)] = \psi^{-1}[(1-q)\psi(n_1) + q\psi(n_2)]$$

holds for

(15) $n_1 = 1$, n_2 arbitrary integer greater than 1, $q \in [0,1]$ arbitrary, if and only if there exist constants $\alpha > 0$, β such that

(16) $$\psi(x) = \alpha\,\phi(x) + \beta \quad \text{for all} \quad x \in [1,\infty[.$$

Proof of the Lemma. The "if" part is obvious. In order to prove the "only if" part, put into (14) $n_1 = 1$, $n_2 > 1$. Denote
$a_1 = \phi(n_1) = \phi(1)$, $a_2 = \phi(n_2) - \phi(n_1) > 0$, $b_1 = \psi(n_1) = \psi(1)$, $b_2 = \psi(n_2) - \psi(n_1) > 0$

Then (14) goes over into

(17) $$\phi^{-1}(a_2 q + a_1) = \psi^{-1}(b_2 q + b_1) \quad (q \in [0,1]).$$

Now denote

$$y = b_2 q + b_1$$

and notice (cf. (15)) that y runs through $[\psi(1), \lim_{n\to\infty} \psi(n)[$ when $q \in [0,1]$, $n_2 = 2,3,\ldots$ (ψ, being increasing, has a finite or infinite limit as $n \to \infty$). So (17) goes over into

$$\psi^{-1}(y) = \phi^{-1}(A_2 y + A_1) \quad (A_2 > 0)$$

or

$$\text{(16)} \qquad \psi(x) = \alpha\phi(x) + \beta \quad \text{for all } x \in [1,\infty[\ ,$$

where $\alpha = 1/A_2 = b_2/a_2 > 0$, q.e.d.

Continuation of the proof of Theorem 1. Denote

$$\psi_m(x) = \phi(x+m) \quad (x \in [1,\infty[;\ m = 1,2,\ldots) \ .$$

Then (12) goes over into

$$\phi^{-1}[(1-q)\phi(n_1) + q\phi(n_2)] = \psi_m^{-1}[(1-q)\psi_m(n_1) + q\psi_m(n_2)]$$

for all

$$q \in [0,1];\ n_1,\ n_2 \ \text{arbitrary integers.}$$

Thus, by the Lemma (the "constants" α, β in (16) will now depend upon m)

$$\text{(18)} \qquad \phi(x+m) = \psi_m(x) = \alpha(m)\phi(x) + \beta(m) \quad (x \in [1,\infty[;\ m = 1,2,\ldots) \ .$$

We distinguish two cases:

(i) $\alpha(m) \equiv 1$. Put then into (18) $x = n$ ($n = 1,2,\ldots$), in order to get

$$\text{(19)} \qquad \phi(m+n) = \phi(n) + \beta(m) \quad \text{for all} \quad m, n = 1,2,\ldots \ .$$

Since the left hand side of (19) is symmetric in m and n, the right hand side has to be symmetric too.

$$\phi(n) + \beta(m) = \phi(m) + \beta(n)$$

and thus (put a constant for n) we have

$$\beta(m) = \phi(m) + c \quad \text{for all} \quad m = 1,2,\dots .$$

This transforms (18) into

(20) $$\phi(x+m) = \phi(x) + \phi(m) + c \qquad (x \in [1,\infty[;\ m = 1,2,\dots) .$$

(ii) If there exists an n_0 such that $\alpha(n_0) \neq 1$, then we derive from (18)

$$\phi(x+m+n) = \alpha(n)\phi(x+m) + \beta(n) = \alpha(m)\alpha(n)\phi(x) + \alpha(n)\beta(m) + \beta(n) .$$

The left hand side is again symmetric in m and n, so the right hand side has to be symmetric too,

$$\alpha(n)\beta(m) + \beta(n) = \alpha(m)\beta(n) + \beta(m)$$

or, with $n = n_0$ $(\alpha(n_0) \neq 1)$, we have

$$\beta(m) = B[\alpha(m) - 1] .$$

Putting this into (18) we get

(21) $$\phi(x+m) = \alpha(m)[\phi(x) + B] - B$$

or, with $x = n$ $(n = 1,2,\dots)$ and again by symmetry,

(22) $$\phi(m+n) + B = \alpha(m)[\phi(n) + B] = \alpha(n)[\phi(m) + B] .$$

By supposition, ϕ is strictly increasing, thus $\phi(n) \not\equiv -B$ and therefore (substitute into (22) $n = n_1$ with $\phi(n_1) \neq -B$)

$$\alpha(m) = a[\phi(m) + B] .$$

Putting this into (21), we finally get

$$\phi(x+m) = a\phi(x)\phi(m) + aB\phi(x) + aB\phi(m) + aB^2 - B . \tag{23}$$

Both (20) and (23) are of the form

$$\phi(x+m) = a\phi(x)\phi(m) + b\phi(x) + b\phi(m) + c \tag{24}$$

with

$$a = 0,\ b = 1 \quad \text{in the case (i)} , \tag{25}$$

and (since ϕ is not constant on $[2,\infty[$)

$$a \neq 0,\ b = aB,\ c = aB^2 - B \quad \text{in the case (ii)} . \tag{26}$$

So (10) goes over into

$$\phi^{-1}(a[p_1\phi(m_1) + p_2\phi(m_2)]\,[q_1\phi(n_1) + q_2\phi(n_2)] + b[p_1\phi(m_1) + p_2\phi(m_2)] + \tag{27}$$

$$+ b[q_1\phi(n_1) + q_2\phi(n_2)] + c) = \phi^{-1}[p_1\phi(m_1) + p_2\phi(m_2)] + \phi^{-1}[q_1\phi(n_1) + q_2\phi(n_2)]$$

with the variables restricted only by (11). If $m_1 = n_1 = 1$ and $m_2, n_2 = 2,3,\ldots,$ then, as p_2 and q_2 run through $[0,1]$,

$$u = p_1\phi(m_1) + p_2\phi(m_2), \quad v = q_1\phi(n_1) + q_2\phi(n_2)$$

assume all values in $[\phi(1), \lim_{n\to\infty} \phi(n)[$ (ϕ being increasing, the finite or infinite limit $\lim_{n\to\infty} \phi(n)$ exists). Therefore (27) goes over into

$$\phi^{-1}(auv+bu+bv+c) = \phi^{-1}(u) + \phi^{-1}(v) \quad \text{for all} \quad u, v \in [\phi(1), \lim_{n\to\infty} \phi(n)[$$

and, with $x = \phi^{-1}(u)$, $y = \phi^{-1}(v)$,

$$(28) \qquad \phi(x+y) = a\phi(x)\phi(y) + b\phi(x) + b\phi(y) + c \quad \text{for all} \quad x, y \in [1,\infty[\ .$$

For the constants in (28) we have one of the two cases (25) or (26). In the case (25), we get that f defined by

$$(29) \qquad f(x) = \phi(x) + c \qquad (x \in [1,\infty[)$$

satisfies the functional equation

$$(30) \qquad f(x+y) = f(x) + f(y) \quad \text{for all} \quad x, y \in [1,\infty[\ .$$

With ϕ also f is increasing, and so, by J. Aczél 1966 and J. Aczél - J. A. Baker - D. Ž. Djoković - Pl. Kannappan - F. Radó 1971, $f(x) = \gamma x$ $(\gamma > 0)$ and

$$(31) \qquad \phi(x) = \gamma x + \delta \quad (\gamma > 0) \quad \text{for all} \quad x \in [1,\infty[\ .$$

In the case (26), we get that g defined by

$$(32) \qquad g(x) = a[\phi(x) + B] \qquad (x \in [1,\infty[;\ a \neq 0)$$

$[g(m) = \alpha(m);\ m = 1,2,\ldots]$ satisfies

$$(33) \qquad g(x+y) = g(x)g(y) \quad \text{for all} \quad x, y \in [1,\infty[\ .$$

From (32) we see that g is strictly monotonic. On the other hand, as (33) shows, if there were an x_0 for which $g(x_0) = 0$ then $g(x_0+y) = 0$ for all $y \in [1,\infty[$ which would contradict the strict monotonicity of g. Thus g is (strictly monotonic and) nowhere zero and, according to the above references,

$$g(x) = D^{tx} \quad (t \neq 0) \quad \text{for all} \quad x \in [1,\infty[$$

and

$$(34) \qquad \phi(x) = \gamma D^{tx} + \delta \quad (\gamma t > 0) \quad \text{for all} \quad x \in [1,\infty[\; .$$

Putting (31) or (34) into (3) we get (4) and (5), respectively, and this concludes the proof of our Theorem 1.

On the other hand, the functions given by (31) and (34) satisfy (8) for all $J > 1$, $K > 1$ [and all m_j, n_k, p_j, q_k $(j = 1,2,\ldots,J;\ k = 1,2,\ldots,K)$ satisfying (6), (9) and (1)], thus the arithmetic and exponential means (4) and (5) are always additive (7).

3. The property (12) or its generalization, both called translativity,

$$(35) \qquad \phi^{-1}\Big[\sum_{k=1}^{K} q_k \phi(n_k+m)\Big] = \phi^{-1}\Big[\sum_{k=1}^{K} q_k \phi(n_k)\Big] + m$$

whenever (1) and (2) are satisfied, is quite important in itself. It serves (cf. J. Aczél 1973) to prove certain uniqueness properties of the so called Shannon and Rényi entropies which are the lower bounds of our mean codeword lengths (4) and (5). We will come back to this later briefly. On the other hand, after allowing non-integer codeword lengths, L. L. Campbell 1966 has deduced (31) and (34) from the (12) translativity alone. Thus, in the case of those generalized codeword lengths, the (12) translativity and the (8) additivity are equivalent. This is not so anymore for the proper positive integer codeword lengths, not even (35) implies (8) or (10) [of course, (8) does imply (35)]. We will give, however, the general solution of the (12) translativity equation and we will show that (35) and (12) are equivalent.

If we have (12) for (13), then we can proceed, as in the proof of Theorem 1, till (24) with (25) or (26). From (24) we get then

$$(36) \quad \phi^{-1}(auv_m+bu+bv_m+c) = \phi^{-1}(u) + \phi^{-1}(v_m) \quad \text{for all} \quad u \in [\phi(1), \lim_{n\to\infty} \phi(n)[,$$

but only for all $v_m = \phi(m)$, $m = 1,2,\ldots$.

However, (24) and (36) imply (35):

$$\phi^{-1}[\sum_{k=1}^{K} q_k\phi(n_k+m)] = \phi^{-1}[a\phi(m)\sum_{k=1}^{K} q_k\phi(n_k) + b\sum_{k=1}^{K} q_k\phi(n_k) + b\phi(m) + c] =$$

$$= \phi^{-1}[\sum_{k=1}^{K} q_k\phi(n_k)] + m .$$

Thus (12) indeed implies (35) and, since (12) is the special case $K = 2$ of (35), the equivalence of these two equations is established.

In order to solve (35) or (12) or, equivalently, (24) in the cases (25) and (26), introduce again the functions f and g defined by (29) and (32), respectively. They will satisfy now the functional equations

$$(37) \qquad f(x+m) = f(x) + f(m) \quad (x \in [1,\infty[;\ m = 1,2,\ldots)$$

and

$$(38) \qquad g(x+m) = g(x)g(m) \quad (x \in [1,\infty[;\ m = 1,2,\ldots) ,$$

respectively. Again ϕ' and thus g can be strictly monotonic only if g is nowhere 0 [$g(x_0) = 0$ would imply $g(x_0+m) = 0$ for all $m = 1,2,\ldots$].

It is easy to construct the general continuous strictly increasing solution of (37):

$$(39) \qquad f(x) = \begin{cases} \text{arbitrary continuous increasing on } [1,2] \text{ but with } f(2) = 2f(1), \\ f(x-k) + kf(1) \quad \text{for} \quad x \in \,]k+1, k+2] \quad (k = 1,2,\ldots) \end{cases}$$

and the general continuous strictly monotonic (increasing, if $a > 0$, decreasing if $a < 0$) solution of (38)

$$(40) \qquad g(x) = \begin{cases} \text{arbitrary strictly monotonic continuous on } [1,2] \text{ but with } g(2) = g(1)^2, \\ g(x-k)g(1)^k \quad \text{for} \quad x \in \,]k+1, k+2] \quad (k = 1,2,\ldots) . \end{cases}$$

So we have proved the following (the "if" part is easily checked).

Theorem 2. The translativity equations (12) and (35) are equivalent. A function ϕ is continuous, strictly increasing and satisfies (12) or (35) if, and only if,

$$\phi(x) = f(x) - c \qquad (x \in [1,\infty[)$$

or

$$\phi(x) = \frac{1}{a} g(x) - B \qquad (x \in [1,\infty[)$$

where $a \neq 0$, B, c are constants and f and g are given by (39) and (40) (g increasing if $a > 0$ and decreasing if $a < 0$).

4. It is well known (F. M. Reza 1961, L. L. Campbell 1965, 1966, J. Aczél 1973) that for all Q and N satisfying (1) and (2), respectively,

$$(41) \qquad L(Q,N;\phi_0) = \sum_{k=1}^{K} q_k n_k \geq - \sum_{k=1}^{K} q_k \log_D q_k, \quad (0 \log 0: = 0)$$

and, <u>for</u> $t > -1,\ t \neq 0$,

$$(42) \qquad L(Q,N;\phi_t) = \frac{1}{t} \log_D \sum_{k=1}^{K} q_k D^{tn_k} \geq \frac{t+1}{t} \log_D \sum_{k=1}^{K} q_k^{1/(t+1)}, \quad (0^{\alpha}: = 0) .$$

The right hand side of (41) is the <u>Shannon entropy</u> while on the right hand side of (42) <u>Rényi entropies</u> [of order $1/(t+1)$] stand.

One advantage of allowing non-integer codeword lengths is (L. L. Campbell 1966), that the lower bounds at the right hand sides of (41) and (42) are actually attained. But even if we restrict ourselves to integer codeword lengths, it is easy to prove (F. M. Reza 1961, L. L. Campbell 1965, J. Aczél 1973) that

$$(43) \qquad L(Q,N^*;\phi_0) = \sum_{k=1}^{K} q_k n_k^* < - \sum_{k=1}^{K} q_k \log_D q_k + 1$$

if

$$(44) \qquad - \log_D q_k \leq n_k^* < - \log_D q_k + 1 \quad (k = 1,2,\ldots,K)$$

and, <u>for all</u> $t \neq -1,\ t \neq 0$,

$$(45) \qquad L(Q,N^*;\phi_t) = \frac{1}{t} \log_D \sum_{k=1}^{K} q_k D^{tn_k^*} < \frac{t+1}{t} \log_D \sum_{k=1}^{K} q_k^{1/(t+1)} + 1 ,$$

if

$$(46) \qquad - \log_D\Big(q_k^{1/(t+1)} \Big/ \sum_{i=1}^{K} q_i^{1/(t+1)}\Big) \leq n_k^* < - \log_D\Big(q_k^{1/(t+1)} \Big/ \sum_{i=1}^{K} q_i^{1/(t+1)}\Big) + 1$$

$$(k = 1,2,\ldots,K) .$$

We can get these from the transitivity of (4) and (5).

As to $t = -1$, it is easy to show that

$$(47) \qquad \lim_{t \to -1} \left(\frac{t+1}{t} \log_D \sum_{k=1}^{K} q_k^{1/(t+1)}\right) = - \log_D \max (q_1,q_2,\ldots,q_K) .$$

(Thus the right hand side of (47) is the Rényi entropy of order ∞.) So, by going over to the limit $t \to -1$ in (42), we get

$$L(Q,N;\phi_{-1}) = - \log_D \sum_{k=1}^{K} q_k D^{-n_k} \geq - \log_D \max (q_1,q_2,\ldots,q_K).$$

More generally, L. L. Campbell has recently proved (communication by correspondance) that for all $t \leq -1$

$$(48) \qquad L(Q,N;\phi_t) = \frac{1}{t} \log_D \sum_{k=1}^{K} q_k D^{tn_k} \geq \frac{1}{t} \log_D \max (q_1,q_2,\ldots,q_K)$$

while (again for $t \leq -1$)

$$(49) \qquad L(Q,N^*;\phi_t) = \frac{1}{t} \log_D \sum_{k=1}^{K} q_k D^{tn^*_k} < \frac{1}{t} \log_D \max (q_1,q_2,\ldots,q_K) + 1$$

if

$$(50) \qquad n^*_{k_0} = 1,\ n^*_k \geq \log_D \frac{D-1}{D(K-1)}\ (k \neq k_0) \quad \text{where} \quad q_{k_0} = \max (q_1,q_2,\ldots,q_K) .$$

(All these $\{n^*_1,n^*_2,\ldots,n^*_k\}$ do also satisfy (2).)

On the right hand sides of (43), (45) and (49), + 1 can be replaced by arbitrarily small $+ \varepsilon > 0$ if we encode sequences of independent messages consecutively.

The minimum or lower bound properties (41), (42) and (48) give interest to the following interpretation of quasiarithmetic mean codeword lengths, cf. L. L. Campbell 1966. The function ϕ in (3) can be understood as cost function, $\phi(n)$ being the cost of using a codeword of length n. It is reasonable to suppose that ϕ is (strictly) increasing on the set of positive integers and then it can always be extended to a function strictly increasing and continuous on $[1,\infty[$. This is suitable because then ϕ^{-1} can be applied on more than a denumerable set.

Now the <u>average cost</u> of encoding the messages $Y = \{\eta_1,\eta_2,\ldots,\eta_k\}$ (probability distribution $Q = \{q_1,q_2,\ldots,q_k\}$) by a distribution $N = \{n_1,n_2,\ldots,n_k\}$ of codeword lengths is

$$C = \sum_{k=1}^{K} q_k \phi(n_k) .$$

A coding problem of some interest is to <u>minimize the cost</u> C by an appropriate choice of the distribution N, subject to the costraint (2). Since $L(Q,N;\phi) = \phi^{-1}(C)$ and ϕ^{-1} is (continuous and) strictly increasing, an equivalent problem is to minimize the mean codeword length $L(Q,N;\phi)$.

There are multiplicative and additive constants contained in the cost functions as given by (31) and (34). (They do not influence the mean codeword lengths (4) and (5).) For calculating the average costs it may be advisable to <u>normalize</u> them. A possible normalization would assign unit cost to encoding a codeword of length 1 and zero cost in the (idealized) case of a codeword of length 0. Then we still have

(51) $$\tilde{\phi}_0(n) = n \qquad (n = 0,1,2,\ldots)$$

but, instead of ϕ_t, we have

(52) $$\tilde{\phi}_t(n) = \frac{D^{tn} - 1}{D^t - 1} \qquad (t \neq 0;\ n = 0,1,2,\ldots)$$

(One of the advantages is that $\tilde{\phi}_0 = \lim_{t\to 0} \tilde{\phi}_t$ while $\phi_0 \neq \lim_{t\to 0} \phi_t$.) The inequalities (41), (42) and (48) show that <u>the average costs cannot be less than</u>

(53) $$-\sum_{k=1}^{K} q_k \log_D q_k \quad (0 \log 0 := 0) \quad \underline{\text{for}} \quad t = 0 ,$$

(54) $$\frac{\left(\sum_{k=1}^{K} q_k^{1/(t+1)}\right)^{t+1} - 1}{D^t - 1} \quad \underline{\text{for}} \quad t \neq 0,\ t > -1 ,$$

<u>and</u>

(55) $$\frac{1 - \max(q_1, q_2, \ldots, q_K)}{1 - D^t} \quad \underline{\text{for}} \quad t \leq -1 ,$$

<u>whenever the cost functions are</u> $\tilde{\phi}$, <u>given by</u>

$$\tilde{\phi}_0(x) = x \quad \underline{\text{and}} \quad \tilde{\phi}_t(x) = \frac{D^{tx} - 1}{D^t - 1} \quad \underline{\text{for}} \quad t \neq 0 \quad (x \in [1,\infty[)$$

[cf. (51), (52)] which, by Theorem 1 and the above, are the normalized forms of the cost functions in all cases of additive mean codeword lengths (8).

The inequalities (44), (46) and (50) show with what N we <u>get near to the lower bounds (53), (54), and (55) of the average costs</u>, respectively.

References

Aczél, J. 1966: Lectures on Functional Equations and Their Applications. Academic Press, New York - London.

Aczél, J. 1973: On Shannon's Inequality, Optimal Coding, and Characterizations of Shannon's and Rényi's Entropies. To be published in Symposia Mathematica, Ist. Naz. Alta Mat., Roma, Academic Press, New York.

Aczél, J., Baker, J. A., Djoković, D. Z., Kannappan, Pl., Radó, F. 1971: Extensions of Certain Homomorphisms of Subsemigroups to Homomorphisms of Groups. Aequationes Math. 6, 263-271.

Campbell, L. L. 1965: A Coding Theorem and Rényi's Entropy. Information and Control 8, 423-429.

Campbell, L. L. 1966: Definition of Entropy by Means of a Coding Problem. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 6, 113-118.

Reza, F. M. 1961: An Introduction to Information Theory. New York - Toronto - London: McGraw-Hill.