

## Entropy: An Inequality

J.-P. ALLOUCHE, M. MENDÈS FRANCE and G. TENENBAUM

*University of Bordeaux I and University of Nancy I*

(Communicated by N. Iwahori)

**Abstract.** We show that the classical Hölder inequality between means of order  $\alpha$ ,  $0 < \alpha \leq 1$ , can be improved on the assumption that the terms are not too often of comparable size. As an application, we derive a general, optimal bound for the entropy of a probability distribution.

### § 1. Introduction.

In two previous articles [1, 2] concerned with the Rudin-Shapiro sequence we stated and used a lemma without reproducing its proof. We believe that this lemma is interesting in its own accord. The object of this paper is to give an extended and generalized version of our lemma — thereby promoted to the status of a theorem. As a consequence we shall obtain an inequality sharpening in some cases the classical Hölder's inequality, and which leads to a new result concerning the entropy of a probability distribution.

### § 2. An inequality.

**THEOREM.** Let  $\lambda$  be a positive real number and let  $\sum_{k=0}^{\infty} x_k$  be a convergent series with non-negative terms. Suppose that

$$(1) \quad \lambda x_n \geq \sum_{k=n+1}^{\infty} x_k, \quad (n=0, 1, 2, \dots).$$

Then, for all  $\alpha$ ,  $0 < \alpha \leq 1$ , we have

$$(2) \quad \sum_{k=0}^{\infty} x_k^\alpha \leq \{(\lambda+1)^\alpha - \lambda^\alpha\}^{-1} \left( \sum_{k=0}^{\infty} x_k \right)^\alpha$$

with equality in the case  $x_k = \{\lambda/(\lambda+1)\}^k$ ,  $k \geq 0$ .

**COROLLARY.** Let  $\{p_k: k \geq 0\}$  be a probability distribution with entropy

$$H := \sum_{k=0}^{\infty} p_k \log \frac{1}{p_k} .$$

Define the function  $F(x) := (x+1)\log(x+1) - x \log x$ ,  $x > 0$ . Then, on the assumption that

$$\lambda := \sup_{n \geq 0} p_n^{-1} \sum_{k=n+1}^{\infty} p_k < \infty$$

we have

$$(3) \quad H \leq F(\lambda) .$$

Furthermore, equality holds in the geometric case  $p_k = (\lambda+1)^{-1} \{\lambda/(\lambda+1)\}^k$ ,  $k \geq 0$ .

Note that (1) is valid in particular for finite sums, that is,

$$(4) \quad \sum_{k=0}^N x_k^\alpha \leq \{(\lambda+1)^\alpha - \lambda^\alpha\}^{-1} \left( \sum_{k=0}^N x_k \right)^\alpha, \quad (0 < \alpha \leq 1),$$

where now

$$\lambda = \lambda_N := \max_{0 \leq n < N} x_n^{-1} \sum_{k=n+1}^N x_k .$$

But Hölder's inequality states that (4) holds with a factor  $(N+1)^{1-\alpha}$  instead of  $\{(\lambda+1)^\alpha - \lambda^\alpha\}^{-1}$ . Thus (4) is certainly sharper when  $\lambda \leq \alpha^{1/(1-\alpha)}(N+1)$ . In this context it is worthwhile to bear in mind that we may always assume that the  $x_n$  are arranged in decreasing order, whence  $\lambda \leq N$ .

Finally, we show that the corollary is a straightforward consequence of the theorem. Indeed, from (2) with  $x_k = p_k$ , we obtain

$$H_\alpha := (1-\alpha)^{-1} \log \left( \sum_{k=0}^{\infty} p_k^\alpha \right) \leq (1-\alpha)^{-1} \log \frac{1}{(\lambda+1)^\alpha - \lambda^\alpha} .$$

The quantity  $H_\alpha$  is Rényi's  $\alpha$ -entropy [3]. We let  $\alpha$  increase to 1 and obtain the required inequality (3).

### §3. Proof of the theorem.

We may plainly normalize the series such that

$$\sum_{k=0}^{\infty} x_k = 1$$

and from now on we shall put  $x_k = p_k$ ,  $k = 0, 1, \dots$ . Our assumption (1)

may be equivalently written as

$$(5) \quad 1 - s_n \leq \beta(1 - s_{n-1}), \quad (n = 0, 1, 2, \dots)$$

where we have set  $s_n := \sum_{0 \leq j \leq n} p_j$ ,  $\beta := \lambda/(\lambda + 1)$ .

Let  $\alpha$  satisfy  $0 < \alpha \leq 1$  and define

$$h = h(\alpha) := \sum_{n=0}^{\infty} p_n^\alpha.$$

The series converges since by (5),  $p_n \leq 1 - s_{n-1} \leq \beta^n$ . Now we put

$$w_n = \begin{cases} \frac{1 - s_n}{1 - s_{n-1}}, & \text{if } s_{n-1} \neq 1 \\ 0, & \text{if } s_{n-1} = 1 \end{cases}$$

and observe that

$$\begin{aligned} h &= \sum_{n=0}^{\infty} (s_n - s_{n-1})^\alpha = \sum_{n=0}^{\infty} (1 - s_{n-1} - (1 - s_n))^\alpha \\ &= (1 - w_0)^\alpha + w_0^\alpha (1 - w_1)^\alpha + w_0^\alpha w_1^\alpha (1 - w_2)^\alpha + \dots \end{aligned}$$

Let  $t := \max h$ , where the maximum is taken over the set of all sequences  $\{w_0, w_1, \dots\}$  in  $[0, \beta]^N$ . Since  $h = h(w_0, w_1, \dots)$  is a continuous function on this compact space, the maximum is actually attained, at  $\{w_0^*, w_1^*, \dots\}$ , say. Then

$$t = (1 - w_0^*)^\alpha + w_0^{*\alpha} h(w_1^*, w_2^*, \dots) \leq (1 - w_0^*)^\alpha + w_0^{*\alpha} t$$

hence

$$t \leq \max_{0 \leq w_0^* \leq \beta} \frac{(1 - w_0^*)^\alpha}{1 - w_0^{*\alpha}} = \frac{(1 - \beta)^\alpha}{1 - \beta^\alpha} = \{(\lambda + 1)^\alpha - \lambda^\alpha\}^{-1},$$

as required.

REMARK. M. Balazard found an alternative proof of this result based on the observation that the function  $x \rightarrow (x + a)^\alpha - x^\alpha$  is strictly decreasing for all  $a > 0$  and  $\alpha$ ,  $0 < \alpha < 1$ . Indeed, taking  $a = x_n$ , we see that the hypothesis (1) implies

$$(\lambda x_n + x_n)^\alpha - (\lambda x_n)^\alpha \leq \left( \sum_{k=n+1}^{\infty} x_k + x_n \right)^\alpha - \left( \sum_{k=n+1}^{\infty} x_k \right)^\alpha$$

or

$$\{(\lambda + 1)^\alpha - \lambda^\alpha\} x_n^\alpha \leq \left( \sum_{k=n}^{\infty} x_k \right)^\alpha - \left( \sum_{k=n+1}^{\infty} x_k \right)^\alpha.$$

Summing this inequality for all  $n=0, 1, 2, \dots$  yields the required estimate (2).

#### §4. Complements.

In this section we give some further precisions to our result. The first is a complete description of the cases of equality in (2).

COMPLEMENT 1. *Suppose that for some  $\alpha, 0 < \alpha < 1$ , we have*

$$(6) \quad \sum_{k=0}^{\infty} p_k^\alpha = \{(\lambda+1)^\alpha - \lambda^\alpha\}^{-1}.$$

*Then  $\{p_k: k \geq 0\}$  is geometric, up to a reordering of the terms. In other words, if we further assume  $p_0 \geq p_1 \geq \dots$ , we have*

$$p_k = (\lambda+1)^{-1} \left\{ \frac{\lambda}{\lambda+1} \right\}^k, \quad (k=0, 1, \dots).$$

This results easily from the proof in Section 3. Indeed (6) implies, in the previous notation, that  $h(w_0, w_1, \dots) = t$ . Hence, as before, we get

$$(7) \quad \begin{aligned} t = h(w_0, w_1, \dots) &= (1-w_0)^\alpha + w_0^\alpha h(w_1, w_2, \dots) \\ &\leq (1-w_0)^\alpha + w_0^\alpha t \end{aligned}$$

which in turn yields

$$t = \frac{(1-\beta)^\alpha}{1-\beta^\alpha} \leq \frac{(1-w_0)^\alpha}{1-w_0^\alpha}.$$

Together with the initial condition  $w_0 \leq \beta$ , this implies that  $w_0 = \beta$  and we infer from (7) that  $h(w_1, w_2, \dots) = t$ . The proof may hence be completed by an obvious iteration.

We note that this result can also be derived from Balazard's proof of our theorem.

Next, we give a generalization of the corollary.

COMPLEMENT 2. *Let  $\{p_k: k \geq 0\}$  be a probability distribution and  $(A_j)_{j \in J}$  be a partition of  $N$ . Put  $P(A_j) = \sum_{k \in A_j} p_k$ ,  $j \in J$ , and suppose that*

$$\Lambda := \sup_{j \in J} \sup_{n \in A_j} p_n^{-1} \sum_{\substack{k > n \\ k \in A_j}} p_k < \infty.$$

*The function  $F$  being defined as in Section 2, we then have*

$$0 \leq \sum_{k=0}^{\infty} p_k \log \frac{1}{p_k} - \sum_{j \in J} P(A_j) \log \frac{1}{P(A_j)} \leq F(\Lambda).$$

Notice that we always have  $\lambda \leq \lambda$ . This result shows that the entropy of a distribution cannot decrease too much by just grouping terms together: "One cannot organize randomness".

We now prove the complement. The left hand inequality simply follows from the convexity of  $x \log(1/x)$ . For the right hand one, we introduce a parameter  $\alpha$ ,  $0 < \alpha < 1$ , and apply (2) to each series  $\sum_{k \in A_j} p_k$ ,  $j \in J$ , noticing that the corresponding conditions of type (1) hold with  $\lambda$  in place of  $\lambda$ . This yields

$$\sum_{k=0}^{\infty} p_k^\alpha = \sum_{j \in J} \sum_{k \in A_j} p_k^\alpha \leq \{(\lambda + 1)^\alpha - \lambda^\alpha\}^{-1} \sum_{j \in J} P(A_j)^\alpha .$$

The required conclusion follows by letting  $\alpha$  tend to 1, as before.

Finally, we mention the continuous analogue of our theorem, which turns out to be very easy.

COMPLEMENT 3. Let  $y = y(x)$  be a function of class  $C^1$  on  $(0, +\infty)$ , with  $y(0) = 0$ ,  $y(+\infty) = 1$ . Suppose that for some  $\lambda > 0$

$$\lambda y' \geq 1 - y \geq 0 .$$

Then for all  $\alpha$ ,  $0 < \alpha < 1$ ,

$$\int_0^\infty (y'(x))^\alpha dx \leq \alpha^{-1} \lambda^{1-\alpha}$$

and

$$\int_0^\infty y'(x) \log \frac{1}{y'(x)} dx \leq 1 + \log \lambda .$$

Furthermore, both inequalities become equalities when  $y(x) = 1 - \exp(-x/\lambda)$ .

PROOF. The hypothesis plainly implies

$$(y')^{\alpha-1} \leq \left( \frac{\lambda}{1-y} \right)^{1-\alpha}$$

hence

$$\begin{aligned} \int_0^\infty (y')^\alpha dx &= \int_0^\infty (y')^{\alpha-1} y' dx \leq \lambda^{1-\alpha} \int_0^\infty (1-y)^{\alpha-1} y' dx \\ &= \lambda^{1-\alpha} \int_0^1 (1-y)^{\alpha-1} dy = \alpha^{-1} \lambda^{1-\alpha} . \end{aligned}$$

The same technique applies for the second case. The last remark is easy to check and left to the reader.

### References

- [1] J.-P. ALLOUCHE and M. MENDES FRANCE, On an extremal property of the Rudin-Shapiro sequence, *Mathematika*, **42** (1985), 33-38.
- [2] M. MENDES FRANCE et G. TENENBAUM, Dimension des courbes planes, papiers pliés et suite de Rudin-Shapiro, *Bull. Soc. Math. France*, **109** (1981), 207-215.
- [3] A. RÉNYI, On measures of entropy and information, *Proceedings of the 4<sup>th</sup> Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 547-561, Univ. of California Press, 1961.

#### *Present Address:*

J.-P. ALLOUCHE AND M. MENDES FRANCE  
DÉPARTEMENT DE MATHÉMATIQUES, CNRS ET UNIVERSITÉ DE BORDEAUX I  
F-33405 TALENCE CEDEX, FRANCE

G. TENENBAUM  
DÉPARTEMENT DE MATHÉMATIQUES, UNIVERSITÉ DE NANCY I  
B.P. 239, 54506 VANDŒUVRE CEDEX, FRANCE