PRAGMATIC ENTROPY FOR FINITELY ADDITIVE PROBABILITIES (*)

by ANDREA SGARRO (in Trieste) (**)

SOMMARIO. - *Si estende la definizione di entropia a distribuzioni di probabilità numerabili finitamente additive. L'impostazione scelta è pragmatica (fa uso di teoremi di codifica). L'entropia di una distribuzione di probabilità finitamente additiva in senso stretto viene posta pari a* $+\infty$, *poiché la corrispondente sorgente stazionaria senza memoria non è comprimibile mediante codici-blocco; in effetti, a parte casi banali, le sorgenti finitamente additive in senso stretto non sono mai comprimibili mediante codici-blocco.*

SUMMARY. - *The definition of entropy is extended to countable finitely additive probability distributions. The approach taken is pragmatic (makes use of coding theorems). The entropy of a properly finitely additive probability distribution is set equal to* $+\infty$, *because the corresponding stationary memoryless source is non-compressible through block-coding; as a matter of fact, apart from trivial cases, properly finitely additive sources are never compressible through block-coding.*

1. Introduction.

In this note we want to extend the definition of entropy to finitely additive probabilities. We shall take a pragmatic view of entropy, in accordance with the original approach taken by Shannon. Roughly speaking, this means that entropy is defined through source coding theorems which tackle the problem of «compressing»

(*) Pervenuto in Redazione il 27 ottobre 1982.

(**) Indirizzo dell'Autore: Istituto di Matematica e Istituto di Elettrotecnica e di Elettronica dell'Università degli Studi - P.le Europa, 1 - 34100 Trieste.

an information source (a stochastic process). (Later entropy was interpreted also as a measure of uncertainty; cf. Hinčin [1], who inaugurated the so-called axiomatic or functional-analysis approach to information measures). More precisely, we shall revisit the classical situation of source block-coding. Our setting requires that the alphabet of the source (tha range of the stochastic process) be at most countable; since in the finite case finite additivity is the same as σ-additivity, we shall use only countable alphabets. (Continuous entropy is a rather intriguing notion; cf., however, [2]). The mathematical analysis will be made extremely simple owing to the following fact: codes are finite sets; since one has to evaluate only probabilities of finite and cofinite sets the algebra of finite and cofinite sets will do; only the probabilities of singletons need to be known.

Our motivation is not only to extend the definition of entropy to the finitely additive case, but also to see how finitely additive probability distributions (p. d. 's) «work» in problems of «applied probability»; we cannot claim, though, that our results are of practical interest to communications engineers. It turns out that sources ruled by non-σ-additive p. d. 's are essentially non-compressible, at least through block-codes. In particular, the entropy of a (strictly) non-σ-additive p.d. will be set equal to $+\infty$, which corresponds to non-compressibility for stationary memoryless sources. We recall that also in the σ-additive case the entropy of a p.d. can be infinite; cf. [3] where a simple criterion for the convergence of the entropy series is given.

The information-theoretic results are contained in section 4. Section 2 is devoted to some preliminaries on finitely additive p. d. 's: they are not given as particular cases of measure-theoretic results since for our algebra the proofs are trivial. Section 3 contains an ad-hoc definition of finitely additive sources; problems relative to a general definition of finitely additive stochastic processes are mentioned. The paper aims to be self-contained; standard references on information theory can be found, e.g., in [4] or in [5]; for the block-coding of sources cf. also, e.g., [6] or [7]. An informative but concise exposition of finitely additive p. d. 's is found in [8].

2. Preliminaries.

All p. d. 's are over the algebra $\mathfrak{F}$ of the finite and cofinite subsets of a countable set, $\Omega = \{\omega_i\}_{i \geq 1}$, say. Since block-codes are finite this is the relevant algebra for us; extra probabilities are not needed.

Definition 1: The *sum* of a p.d. μ is the number

$$s = s(\mu) \doteq \Sigma\, \mu(\omega_i).$$

(We denote by ω_i also the corresponding singleton; unspecified summations are over all values of the index). Obviously $0 \leqq s(\mu) \leqq 1$.

We give some simple properties of countable p.d.'s.

Lemma 1: Any sequence $\{\mu_i\}_{i \geqq 1}$, $\mu_i \geqq 0$, $\Sigma \mu_i \leqq 1$ identifies a unique p.d. μ over $\mathfrak{F}$ such that $\mu(\omega_i) = \mu_i$.

Proof: If $\mathcal{A}$ is finite set $\mu(\mathcal{A}) = \sum_{\omega_i \in \mathcal{A}} \mu(\omega_i)$; if $\mathcal{A}$ is cofinite set $\mu(\mathcal{A}) = 1 - \mu(\Omega - \mathcal{A})$. Clearly μ is the required p.d. ■

Remark to lemma 1: It is well known that a sequence like that in lemma 1 with $\Sigma \mu_i = 1$ identifies a usual σ-additive p.d. over the algebra of all subsets; over the latter algebra, instead, the numbers μ_i are not enough to identify a finitely additive p.d. Restricting attention to $\mathfrak{F}$, therefore, amounts to consider only that «part» of the p.d. which is completely described by the probabilities of singletons.

Lemma 2: $s(\mu) = \sup_{\mathcal{A} \text{ finite}} \mu(\mathcal{A}) = 1 - \inf_{\mathcal{A} \text{ cofinite}} \mu(\mathcal{A})$

Proof: Obvious. ■

Lemma 3: $s(\mu) = 1$ iff μ is σ-additive (over $\mathfrak{F}$).

Proof: The if is obvious. The only if follows from lemma 1 and from the first statement in the remark following lemma 1. ■

Let U denote the «uniform» p.d. over $\mathfrak{F}$, which is 0 on finite sets and 1 on cofinite sets.

Lemma 4: $\mu = \alpha Q + (1 - \alpha) U$, convex combination, where Q is a σ-additive p.d. If $\mu \neq U$, this decomposition is unique over $\mathfrak{F}$. In any case $\alpha = s(\mu)$.

Proof: If $\mu = U$ the lemma is obvious with $\alpha = s(\mu) = 0$, Q arbitrary. In any case, however, one must have

$$\mu(\omega_i) = \alpha Q(\omega_i) + (1 - \alpha) U(\omega_i) = \alpha Q(\omega_i)$$

and therefore, for $\alpha > 0$, $Q(\omega_i) = \frac{1}{\alpha} \mu(\omega_i)$. Summing over all i's one obtains $\alpha = s(\mu)$. The rest of the lemma follows easily from lemma 1. ■

Notice that for a properly non-σ-additive p.d. $\mu(\mathcal{A}) = 1$ implies $\mathcal{A}$ cofinite. We are now able to compute probabilities of countable unions. Assume $\mathcal{A} = \cup \mathcal{A}_i$, $i \geqq 1$, $\mathcal{A}_i$ disjoint sets. The only nontrivial case is when infinitely many $\mathcal{A}_i$ are not void; then no $\mathcal{A}_i$ is cofinite (two cofinite sets are never disjoint) and $\mathcal{A}$ is cofinite; then

$$\mu(\mathcal{A}) = \mu(\bigcup_{i \geqq 1} \mathcal{A}_i) = \sum_{i \geqq 1} \mu(\mathcal{A}_i) + 1 - s(\mu)$$

(use the decomposition in lemma 4).

3. Sources.

We shall use a rather ad-hoc definition of stochastic process, which bypasses random variables and Kolmogorov's extension theorem. Since in information theory this convenient approach is rather familiar, we shall use the information-theoretic term source, which may sound less committal to some.

Let μ be a p.d. over $\Omega = \{\omega_i\}_{i \geq 1}$. Let $\{\Gamma_n\}_{n \geq 1}$ be a sequence of doubly infinite stochastic matrices with rows labelled in Ω^n and columns labelled in Ω; a row of Γ_n is a sequence of non-negative real numbers whose sum is at most one; $n \geq 1$. The Γ_n element of position $(x^{(n)}, \omega_i)$ will be denoted by $\gamma_n(x^{(n)}, \omega_i)$. We are now able to define a sequence of p.d.'s $\mu^{(n)}$ over Ω^n in the following way:

$$(*) \qquad \mu^{(n+1)}(x^{(n)}\omega_i) = \mu^{(n)}(x^{(n)})\,\gamma_n(x^{(n)}, \omega_i), \;\; \mu^{(1)} = \mu.$$

Note that the rows $\{\gamma_n(x^{(n)}, \omega_i)\}_{i \geq 1}$ of Γ_n for which $\mu^{(n)}(x^{(n)}) = 0$ are irrelevant.

Definition 2: Let μ and $\{\Gamma_n\}_{n \geq 1}$ be as above. The sequence of p.d.'s $\{\mu^{(n)}\}_{n \geq 1}$ given by $(*)$ is called a *source* over the *alphabet* Ω. If $\gamma_n(x^{(n)}, \omega_i)$ does not depend on $x^{(n)}$ the source is said to be *memoryless*; if $\gamma_n(x^{(n)}, \omega_i) = \mu(\omega_i)$ the source is said to be *stationary memoryless.*

Clearly a memoryless source is identified by a sequence of p.d.'s over $\Omega, \{\mu_n\}_{n \geq 1}$; a stationary memoryless source is identified by μ.

Set $s^{(n)} = s(\mu^{(n)})$; in the memoryless case set $s_n = s(\mu_n), n \geq 1$; in the stationary memoryless case set $s = s(\mu)$.

Lemma 5: In the memoryless case $s^{(n)} = \prod_{i=1}^{n} s_i$; in the stationary memoryless case $s^{(n)} = s^n$.

Proof: Use, e.g., lemma 2. ■

Some comments are needed. The p.d.'s $\mu^{(n)}$ are not given over product algebras, but over the algebras $\mathfrak{F}^{(n)}$ made up by the finite and cofinite sets of Ω^n. In the ordinary σ-additive case a «short description» like ours, which makes use only of μ and the Γ_n, is enough to determine the probability of any subset of Ω^n. This is not so in the finitely additive case. In particular there is no way to compute marginal p.d.'s since sets like $\{\omega_1\} \times \Omega$ or $\Omega \times \{\omega_1\} \subset \Omega^2$, say, are neither finite nor cofinite; it does not even make much sense to say that $\mu^{(n)}$ is an «initial» marginal for $\mu^{(n+1)}$. As a matter of fact our definitions of memoryless sources and of stationary memoryless sources (or, rather, our choice of the terms «memoryless» and «stationary memoryless») are questionable; events which «do not begin at initial time» are ruled out; in particular we are unable to define stationarity by itself. To discuss properly memory and stationarity, therefore, larger product algebras are needed: in this more general context our definitions would be better suited to the lower status of necessary conditions.

4. Entropy.

Let $\mathcal{S}$ be a source as in def. 2, $\mathcal{S} = \{\mu^{(n)}\}_{n\geq 1}$. An (optimal block-) R-*code* of length n, $R > 0$, is a subset of Ω^n of maximal probability subject to the constraint that its cardinality is not greater than $\exp(nR)$. Let $\mathcal{A}^{(n)}$ be an R-code; its error probability $P^{(n)}_e = P^{(n)}_e(R)$ is defined as

$$P^{(n)}_e(R) = \mu^{(n)}(\Omega^n - \mathcal{A}^{(n)}) = 1 - \mu^{(n)}(\mathcal{A}^{(n)})$$

Four quantities stand as candidates for the entropy of $\mathcal{S}$, $H(\mathcal{S})$, $(\inf \phi = +\infty)$:

$$\inf\{R\colon \liminf_n P_e^{(n)}(R) < 1\}, \quad \inf\{R\colon \limsup_n P_e^{(n)}(R) < 1\},$$

$$(**)\ \inf\{R\colon \liminf_n P_e^{(n)}(R) = 0\},$$

$$\inf\{R\colon \limsup_n P_e^{(n)}(R) = \lim P_e^{(n)}(R) = 0\}.$$

It is well-known that in many important cases these four infima coincide, leaving no room for doubts.

Definition 3: In the four infima (**) coincide, any of them is called the *entropy* of source $\mathcal{S}, H(\mathcal{S})$.

Before proceeding we indulge in some «functional» heuristics. Let us go back to the decomposition which appears in lemma 4, and let us assume that $\alpha \neq 1$, so that U actually appears in the decomposition (μ is properly non-σ-additive). The «usual» entropy $H(\mu)$ is concave, so that one would expect

$$H(\mu) \geq \alpha H(Q) + (1-\alpha) H(U).$$

On the other hand U can be seen as the «limit» of uniform distributions U_n over $\{\omega_1, \omega_2, \ldots, \omega_n\}$ as n goes to infinity. Since the «usual» entropy is continuous this would give

$$H(U) = \lim_n H(U_n) = \lim_n \log n = +\infty,$$

and therefore also $H(\mu) = +\infty$ for any μ with $s(\mu) < 1$. Instead we shall pursue the pragmatic approach to entropy. Nevertheless it will turn out that also this approach gives the same result.

An extremely rough bound on $P^{(n)}_e(R)$ can be given using only the fact that a code is a finite set. As a matter of fact $\mu^{(n)}(\mathcal{A}^{(n)}) \leq s^{(n)}$ (lemma 2) gives

(***) $$P_e^{(n)}(R) \geqq 1 - s^{(n)}$$

This is enough to state simple conditions ensuring that (some of) the four infima (**) are $+\infty$. In particular:

Proposition 1: For a stationary memoryless source $\mathcal{S} = \{\mu^{(n)}\}_{n\geqq 1}$, $\mu^{(n)} = \mu$, with $s(\mu) < 1$ $H(\mathcal{S}) = +\infty$.

Proof: Use lemma 5. ■

Roughly speaking this means that stationary memoryless sources are intractable from the point of view of compressibility if they are ruled by a properly non-σ-additive p.d. (We recall that such «intractable» sources exist even in the σ-additive case; see the introduction). Proposition 1 entitles us to give the following definition:

Definition 4: Let μ be a p.d. over Ω with $s(\mu) < 1$; then the entropy of $\mu, H(\mu)$, is set equal to $+\infty$.

(***) implies also:

Proposition 2: If $H(\mathcal{S}) < +\infty$, then $\lim\limits_n s^{(n)} = 1$.

Proof: Obvious. ■

Therefore «asymptotic» σ-additivity is required for the entropy to be finite: in this sense no «essentially» non-σ-additive source is compressible.

We shall give a more detailed analysis of the error probability and make it amenable to the error probability of an ordinary σ-additive source. We need some observations: 1st: To obtain an R-code one can list the elements of Ω^n in order of non-increasing probability and then take as many as possible starting from the beginning. 2nd: Set $\mu^{(n)} = \alpha^{(n)} Q^{(n)} + (1 - \alpha^{(n)}) U$, $Q^{(n)}$ σ-additive, as in lemma 4. Since $\mu^{(n)}(x^{(n)}) = \alpha^{(n)} Q^{(n)}(x^{(n)})$, if $\alpha^{(n)} \neq 0$ the order of non-increasing probability is the same for $\mu^{(n)}$ and $Q^{(n)}$. Therefore an R-code for the dummy σ-additive source $\tilde{\mathcal{S}} = \{Q^{(n)}\}_{n\geqq 1}$ is also an R-code for the original source $\mathcal{S} = \{\mu^{(n)}\}_{n\geqq 1}$; this is also (trivially) true for $\alpha^{(n)} = 0$. (To be fastidious, the $Q^{(n)}$ should be prolonged over the σ-algebras of all subsets to get an «ordinary» source; extra probabilities, however, are irrelevant).

If $\mathcal{C}^{(n)}$ is an R-code, one has:

$$P_e^{(n)}(R) = \mu^{(n)}(\Omega^n - \mathcal{C}^{(n)}) = 1 - \alpha^{(n)} + \alpha^{(n)} Q^{(n)}(\Omega^n - \mathcal{C}^{(n)}) =$$

$$= 1 - \alpha^{(n)} + \alpha^{(n)} \tilde{P}_e^{(n)}(R)$$

where $\tilde{P}_e^{(n)}(R)$ is the error probability corresponding to the ordinary source $\tilde{\mathbb{S}}$. Thus the analysis of the true error probability $P^{(n)}_{e}(R)$ is amenable to the analysis of the error probability for the σ-additive source $\tilde{\mathbb{S}}$.

Remark: In section 3 we have argued that stronger definitions of memoryless and stationary memoryless sources are needed. Proposition 1 would then hold true a fortiori; in particular definition 4 would still be heuristically sound.

REFERENCES

[1] A. I. KHINCIN, *Mathematical foundations of information theory,* (1957) Dover, New York.

[2] K. OZEKI, *A new definition of the entropy of general probability distributions based on the non-standard analysis,* Information and control, 47, pp. 94-106 (1980).

[3] A. D. WYNER, *An upper bound on the entropy series,* Information and control, 20, pp. 176-181 (1972).

[4] I. CSISZÁR and J. KÖRNER, *Information theory. Coding theorems for discrete memoryless systems,* (1981) Academic Press.

[5] G. LONGO, *Teoria dell'informazione,* (1981), Boringhieri.

[6] G. LONGO and A. SGARRO, *The source coding theorem revisited: a combinatorial approach,* IEEE transactions on information theory, IT-25, 5 (1979) pp. 544-548.

[7] L. DAVISSON, G. LONGO and A. SGARRO, *The error exponent for the noiseless encoding of finite ergodic Markov sources,* IEEE transactions on information theory, IT-27, 4 (1981) pp. 431-438.

[8] R. SCOZZAFAVA, *Probabilità σ-additive e non,* Bollettino dell'UMI, serie VI, vol. I-A, n. 1 (1982), pp. 1-33.