AN ELEMENTARY PROOF OF THE AEP OF INFORMATION THEORY[1]

By A. J. Thomasian

University of California, Berkeley

1. Summary. Properties of the sequence of random variables $-(1/n) \log p$ are obtained for an arbitrary, not necessarily ergodic or stationary, information source. These permit an elementary combinatorial proof of the AEP (asymptotic equipartition property).

2. Definitions and introduction. Let A be a set of $r \geqq 2$ symbols and let $A^{(n)}$ be the set of n-tuples from A. We call A an alphabet, and an element of $A^{(n)}$ a message of length n. Let A^I be the set of infinite sequences $(y_1, y_2, \cdots)$ where each $y_i \in A$, and let P be a probability distribution on the σ-field of subsets of A^I determined by the cylinder sets. We call (A^I, P) an information source and define a sequence of nonnegative random variables X_n by

$$\begin{aligned} X_n(y_1, y_2, \cdots) \\ &= -n^{-1} \log P[Y_1 = x_1, \cdots, Y_n = y_n] \quad \text{if } P[Y_1 = y_1, \cdots, Y_n = y_n] > 0 \\ &= 0 \quad \text{if } P[Y_1 = y_1, \cdots, Y_n = y_n] = 0 \end{aligned}$$

where all our logarithms are to the base 2. In extending Shannon's work [5], McMillan [4] introduced the definition that a source has the AEP if X_n converges in probability to a constant. For a stationary ergodic process, McMillan [4] proved that X_n converged to the constant given in Section 4 in L^1 mean and in probability; while Breiman [1] obtained convergence with probability one. Both proofs use an ergodic theorem and martingales. The proofs of Feinstein [2] and Khinchin [3] follow McMillan.

For any integer n and any number β, define $D_n(\beta)$ to be the largest probability of any subset of $A^{(n)}$ which has at most $2^{\beta n}$ elements. In Section 3 we obtain relations between $P[X_n \leqq \beta]$, $D_n(\beta)$, and EX_n for an arbitrary source. In Section 4 we restrict ourselves to stationary ergodic sources and use $D_n(\beta)$ and Theorem 3 to prove the AEP. Except for two simple properties of entropy, Shannon [5] or Khinchin [3], p. 4 and p. 6, the paper is self-contained.

Henceforth we will consider a fixed source (A^I, P) and its associated r, X_n, $D_n(\beta)$.

3. Relations between $P(X_n \leqq \beta)$, $D_n(\beta)$, and EX_n.

Lemma 1. *For all* $\epsilon > 0$, β

$$P[X_n \leqq \beta] \leqq D_n(\beta) \leqq P[X_n \leqq \beta + \epsilon] + 2^{-\epsilon n}.$$

Received August 24, 1959.

[1] This research was supported by the Office of Naval Research under Contract Nonr-222(53).

PROOF. Let B be the subset of elements of $A^{(n)}$ which have positive probability and which belong to $[X_n \leqq \beta]$, and let M be the number of elements in B. Any point in B has probability $\geqq 2^{-\beta n}$ so that $1 \geqq P[X_n \leqq \beta] \geqq M2^{-\beta n}$. Thus $M \leqq 2^{\beta n}$ so that $D_n(\beta) \geqq P(B)$ and the left-hand inequality is proved.

Let $F \subseteq A^{(n)}$ have at most $2^{\beta n}$ elements and satisfy $P(F) = D_n(\beta)$. Then

$$D_n(\beta) = P(F) = P(F \cap [X_n \leqq \beta + \epsilon]) + P(F \cap [X_n > \beta + \epsilon]) \leqq P[X_n \leqq \beta + \epsilon] + 2^{\beta n}2^{-(\beta+\epsilon)n},$$

and the right-hand inequality is proved.

LEMMA 2. *Let β_n be any sequence of numbers. Then*

(a) $D_n(\beta_n + \epsilon) \to 1$ *for all* $\epsilon > 0$ *if and only if* $P[X_n \leqq \beta_n + \epsilon] \to 1$ *for all* $\epsilon > 0$.

(b) $D_n(\beta_n - \epsilon) \to 0$ *for all* $\epsilon > 0$ *if and only if* $P[X_n \leqq \beta_n - \epsilon] \to 0$ *for all* $\epsilon > 0$.

PROOF. Immediate from Lemma 1.

We pause for a moment to ask an incidental question. If there is a number β such that $X_n \xrightarrow{P} \beta$ (X_n converges in probability to β) and if for each n we select β_n so that $D_n(\beta_n)$ is approximately .8, then, must $\beta_n \to \beta$? The answer is yes by Theorem 1, which generalizes similar theorems proved by Shannon [5], Theorem 4, and Khinchin [3], Theorem 3, p. 20.

THEOREM 1. *If α, β, β_n are numbers such that $X_n \xrightarrow{P} \beta$, $0 < \alpha < 1$, and $\alpha < D_n(\beta_n) < 1 - \alpha$ for all n, then $\beta_n \to \beta$.*

PROOF. If β_n does not converge to β, then there is an $\epsilon > 0$ and a subsequence $\beta_{n'}$ such that either $\beta_{n'} \geqq \beta + \epsilon$ for all n', or $\beta_{n'} \leqq \beta - \epsilon$ for all n'. In either case the assumption $\alpha < D_n(\beta_n) < 1 - \alpha$ is contradicted by Lemma 2 with β_n replaced by β, so that Theorem 1 is proved.

Theorem 2 and Lemma 4 show that to some extent the random variables X_n enjoy some of the properties of a sequence of uniformly bounded random variables. The proofs are based on

LEMMA 3. *For any numbers $\epsilon > 0$, $\delta > 0$, $\beta \geqq 0$*

(a) $$\delta P[X_n < \beta - \delta] \leqq (\beta - EX_n) + \epsilon + (\log r)P[X_n > \beta + \epsilon] - n^{-1}P[X_n > \beta + \epsilon] \log P[X_n > \beta + \epsilon]$$

(b) $$\epsilon P[X_n > \beta + \epsilon] \leqq (EX_n - \beta) + \delta + (\beta - \delta)P[X_n < \beta - \delta].$$

PROOF. We first prove (a). $EX_n = \int_{[X_n<\beta-\delta]} X_n\,dP + \int_{[\beta-\delta\leqq X_n\leqq\beta+\epsilon]} X_n\,dP + \int_{[X_n>\beta+\epsilon]} X_n\,dP$. Thus

$$EX_n \leqq (\beta - \delta)P[X_n < \beta - \delta] + (\beta + \epsilon)(1 - P[X_n < \beta - \delta]) + \int_{[X_n>\beta+\epsilon]} X_n\,dP \leqq (\beta + \epsilon) - (\delta + \epsilon)P[X_n < \beta - \delta] + \int_{[X_n>\beta+\epsilon]} X_n\,dP,$$

so that we need only show that $\int_{[X_n>\beta+\epsilon]} X_n\, dP \leqq (\log r)p - 1/n\, p \log p$ where $p = P[X_n > \beta + \epsilon]$. Now we recall, Khinchin [3], p. 4, that if $p_i > 0$ and $\sum_1^k p_i = p$, then

$$-\sum_{i=1}^{k} \frac{p_i}{p} \log \frac{p_i}{p} \leqq \log k,$$

so that

$$-\sum_{i=1}^{k} p_i \log p_i \leqq p \log k - p \log p.$$

Since $A^{(n)}$ has r^n points we see that $\int_{[X_n>\beta+\epsilon]} X_n\, dP \leqq 1/n[p \log r^n - p \log p]$ and the proof of part (a) is completed.

To prove part (b) we start from the same initial decomposition of EX_n as in part (a) and obtain

$$\begin{aligned} EX_n &\geqq (\beta - \delta)(1 - P[X_n < \beta - \delta] - P[X_n > \beta + \epsilon]) + (\beta + \epsilon)P[X_n > \beta + \epsilon] \\ &\geqq (\beta - \delta) - (\beta - \delta)P[X_n < \beta - \delta] + (\delta + \epsilon)P[X_n > \beta + \epsilon], \end{aligned}$$

so that part (b) is proved.

THEOREM 2. *If for some β we have $X_n \xrightarrow{P} \beta$ then $EX_n \to \beta$.*

PROOF. Immediate from Lemma 3.

The next result in a sense permits us to eliminate half of our task whenever we try to prove that $X_n - EX_n \xrightarrow{P} 0$.

LEMMA 4. $P[X_n \leqq EX_n + \epsilon] \to 1$ *for all* $\epsilon > 0$ *if and only if*

$$P[X_n \leqq EX_n - \epsilon] \to 0$$

for all $\epsilon > 0$.

PROOF. Immediate from Lemma 3 when β is replaced by EX_n.

THEOREM 3. *If EX_n converges to some number β and any one of*

$$P[X_n \leqq \beta + \epsilon] \to 1, \qquad P[X_n \leqq \beta - \epsilon] \to 0, \qquad D_n(\beta + \epsilon) \to 1, \qquad D_n(\beta - \epsilon) \to 0$$

is true for all $\epsilon > 0$, then $X_n \xrightarrow{P} \beta$.

PROOF. Immediate from Lemmas 2 and 4.

4. Proof of the AEP. Henceforth we will consider only stationary sources. Thus we assume, for all $k \geqq 1$ and for all $(y_1, \cdots, y_k) \in A^{(k)}$, that

$$P[Y_{j+1} = y_1, \cdots, Y_{j+k} = y_k]$$

is independent of $j \geqq 0$. For any $m \geqq 2$, $I \in A^{(m-1)}$, $j \in A$ we mean by (I, j) that element of $A^{(m)}$ whose first $m - 1$ coordinates agree with I and whose last coordinate is j; and we define $q_I = P(I)$ and $q_{Ij} = P(I, j)/P(I)$ for $P(I) > 0$. Let $H_m = -\sum q_I q_{Ij} \log q_{Ij}$, where the sum is over all (I, j) with $q_I > 0$.

Clearly $H_m \geqq 0$ and it is well known, Khinchin [3], p. 6, that $H_m \leqq \log r$ is non-increasing so that

$$EX_n = \frac{1}{n} \sum_{j=1}^{n} H_j \to H = \lim H_m .$$

If X_n converges in probability to some constant, then we know from Theorem 2 that this constant must be H.

For a given $m \geqq 2$, and $I \varepsilon A^{(m-1)}$, $j \varepsilon A$, n we define the random variable $N_{Ij}^n = N_{Ij}^n(y_1, y_2, \cdots)$ as the number of integers i with $1 \leqq i \leqq n - m + 1$ such that $(y_i, y_{i+1}, \cdots, y_{i+m-1}) = (I, j)$. We will call a stationary source ergodic if for all m, I, j

$$N_{Ij}^n/n \xrightarrow{P} q_I q_{Ij} .$$

This definition of ergodic is intuitively appealing and it is precisely this property which we will use in our proof of the AEP. It is easy to show, Khinchin [3], p. 49, that our definition of ergodic is equivalent to the usual one.

THEOREM 4. $X_n \xrightarrow{P} H$ *for any stationary ergodic source.*

PROOF. It is clear from Theorem 3 that it is sufficient to prove that for all m, $\epsilon > 0$ we have $D_n(H_m + \epsilon) \to 1$. We will do this by exhibiting, for every m, $\epsilon > 0$, a sequence of sets $B_n \subset A^{(n)}$ with $P(B_n) \to 1$ and

$$M(B_n) \leqq 2^{(H_m+\epsilon)n} \qquad \text{for all large } n$$

where $M(B_n)$ is the number of elements in B_n.

Let m, $\epsilon > 0$ be given, and for arbitrary $\delta > 0$ define B_n as the set of $(y_1, y_2, \cdots)$ such that

$$\left| \frac{N_{Ij}^n(y_1, y_2, \cdots)}{n} - q_I q_{Ij} \right| \leqq \delta \quad \text{for all } I, j \quad \text{with } q_I q_{Ij} > 0$$

and

$$N_{Ij}^n(y_1, y_2, \cdots) = 0 \quad \text{for all } I, j \quad \text{with } q_I = 0 \text{ or } q_{Ij} = 0.$$

Clearly $B_n \subset A^{(n)}$ and $P(B_n) \to 1$, so that we need only bound $M(B_n)$ appropriately, to complete the proof. We now use the q_{Ij}, q_I to define a new stochastic process, with probability distribution Q on A^I, which is to be a multiple Markov chain. Thus we start our Q process off with q_I as the initial distribution and use q_{Ij} for our transition probabilities. For the Q process, for any $n \geqq m$ and any

$$(y_1, \cdots, y_n) \varepsilon A^{(n)},$$

we have

$$Q[Y_n = y_n \mid Y_1 = y_1, \cdots, Y_{n-1} = y_{n-1}]$$
$$= Q[Y_n = y_n \mid Y_{n-m+1} = y_{n-m+1}, \cdots Y_{n-1} = y_{n-1}];$$

that is, the conditional probabilities of future states depend only on the $m - 1$ past states. Now for any $(y_1, \cdots, y_n) \varepsilon B_n$ we have

$$Q(y_1, \cdots, y_n) = q_{I'} \prod (q_{Ij})^{N^n_{Ij}(y_1, \cdots, y_n)},$$

where $q_{I'} > 0$ and the product is over all (I, j) with $q_I q_{Ij} > 0$. Since

$$(y_1, \cdots, y_n) \varepsilon B_n$$

we have

$$N^n_{Ij}(y_1, \cdots, y_n) \leqq (q_I q_{Ij} + \delta)n,$$

so that

$$Q(y_1, \cdots, y_n) \geqq [(q_{I'})^{1/n} (\prod q_{Ij})^{\delta} \prod (q_{Ij})^{q_I q_{J}}]^n \geqq 2^{-(H_m + \epsilon)n},$$

where the last inequality is obtained for δ small enough and n large enough so that $(q_{I'})^{1/n} (\prod q_{Ij})^{\delta} \geqq 2^{-\epsilon}$. Under these conditions

$$1 \geqq Q(B_n) \geqq 2^{-(H_m+\epsilon)n} M(B_n),$$

and the proof is completed.

REFERENCES

[1] L. Breiman, "The individual ergodic theorem of information theory," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 809–811.

[2] A. Feinstein, *Foundations of Information Theory*, McGraw-Hill, New York, 1958.

[3] A. I. Khinchin, *Mathematical Foundations of Information Theory*, Dover Publications, 1957.

[4] B. McMillan, "The basic theorems of information theory," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 196–219.

[5] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, Vol. 27 (1948), pp. 379–423, and pp. 623–656.