SEPARATUM

Studia Scientiarum Mathematicarum Hungarica

Tomus II Fasciculi 1—2

G. KATONA and G. TUSNÁDY

THE PRINCIPLE OF CONSERVATION OF ENTROPY IN A NOISELESS CHANNEL

1967

Studia Scientiarum Mathematicarum Hungarica 2 (1967) 29—35.

THE PRINCIPLE OF CONSERVATION OF ENTROPY IN A NOISELESS CHANNEL

by

G. KATONA and G. TUSNÁDY

Introduction

The main aim of this paper is to formulate precisely and to prove the following statement:

If we have an information source, more precisely, a sequence of random variables $\xi_1, \xi_2, \dots$ with entropy $H(\mathcal{X})$ and we code this sequence in a uniquely decodable manner, the obtained sequence $\mathcal{Y}$ has the entropy

$$H(\mathcal{Y}) = \frac{H(\mathcal{X})}{L}, \tag{1}$$

where L is the average length of the codes.

The intuitive meaning of (1) is clear: it expresses the principle of conservation of information, when the coding is uniquely decodable (and no noise is present). In spite of this according to our best knowledge (1) has not been proved in full generality up to now.

If we have finite number of code signals $y_1, \dots, y_m$, the maximum of $H(\mathcal{Y})$ is $\log m$, where log denotes the logarithm with respect to the base 2. It follows from (1) that

$$\frac{H(\mathcal{X})}{\log m} \leqq L.$$

This is a well known theorem of SHANNON, and according to our best knowledge only this consequence of (1) was proved (e. g. [1]).

In the case when the coding is not necessarily uniquely decodable instead of (1) we prove the inequality

$$H(\mathcal{Y}) \leqq \frac{H(\mathcal{X})}{L},$$

which has also an intuitive meaning.

Precise Formulation

Let $X = \{x_1, \dots, x_n\}$ be the set of possible signals (the alphabet) of the information source, and let X^∞ be the set of all infinite sequences formed from the letters $x_1, \dots, x_n$. If $1 \leqq i_1 \leqq n, \dots, 1 \leqq i_k \leqq n$, we denote by $[x_{i_1}, \dots, x_{i_k}]$ the set of all sequences having $x_{i_1}, \dots, x_{i_k}$ on the first k places. We call such subsets of X^∞ cylinder sets. Let $\mathfrak{A}_X$ denote the σ-field generated by the cylinder sets. The measure space $\mathcal{X} = (X^\infty, \mathfrak{A}_X, p_X)$

is called an information source, if p_X is a probability measure on $\mathfrak{A}_X$. This space defines an other space on the sequences of length k: $\mathscr{X}^k=(X^k, \mathfrak{A}_X^k, p_X^k)$, where X^k is the space of the sequences $(x_{i_1}, \ldots, x_{i_k})$. The average information contained in the first k signals of the information source is

$$H(\mathscr{X}^k) = -\sum_{\substack{1\leqq i_1\leqq n\\ \vdots\\ 1\leqq i_k\leqq n}} p_X^k(x_{i_1}, \ldots, x_{i_k}) \log p_X^k(x_{i_1}, \ldots, x_{i_k}) =$$

$$= -\sum_{\substack{1\leqq i_1\leqq n\\ \vdots\\ 1\leqq i_k\leqq n}} p_X[x_{i_1}, \ldots, x_{i_k}] \log p_X[x_{i_1}, \ldots, x_{i_k}].$$

Finally, the definition of the entropy of $\mathscr{X}$ is

$$H(\mathscr{X}) = \lim_{k\to\infty} \frac{H(\mathscr{X}^k)}{k}$$

(the average information content of one signal), if this limit exists.

Consider now the definition of the coding. Let $Y=\{y_1, \ldots, y_m\}$ be the set of possible code signals. Let $c(x_i)$ $(1\leqq i\leqq n)$ be a finite, non empty sequence formed from elements of Y. We call this function coding. Thus we may associate to every $(x_{i_1}, \ldots, x_{i_k})$, resp. $(x_{j_1}, x_{j_2}, \ldots)$ a sequence of y_i's. Let us write successively the codes $c(x_{i_1}), \ldots, \ldots, c(x_{i_k})$, resp. $c(x_{j_1}), c(x_{j_2}), \ldots$. Denote by $c(x_{i_1}, \ldots, x_{i_k})$, resp. $d(x_{j_1}, x_{j_2}, \ldots)$ the resulting sequence. Let Y^∞ be the set of all infinite y-sequences. Thus the function $d(x_{j_1}, x_{j_2}, \ldots)$ transforms the set X^∞ into a subset Y^* of Y^∞. Let $\mathfrak{A}_Y$ be the σ-field in Y^* generated by the mapping $d(x_{j_1}, x_{j_2}, \ldots)$ of $\mathfrak{A}_X$ and let us define the measure P_Y on $\mathfrak{A}_Y$ by putting:

$$p_Y(A)=p_X(d^{-1}(A)) \quad A\in\mathfrak{A}_Y$$

where $d^{-1}(A)$ denotes the inverse image of A. $\mathscr{Y}=(Y^*, \mathfrak{A}_Y, p_Y)$ is the space of coded sequences. As above $[y_{i_1}, \ldots, y_{i_k}]$ denotes the cylinder set consisting of all sequences in Y^∞ of which the first k terms are $y_{i_1}, \ldots, y_{i_k}$.

LEMMA 1. $[y_{i_1}, \ldots, y_{i_k}]\cap Y^*\in\mathfrak{A}_Y$.

PROOF. We have to prove that the set of all sequences $(x_{j_1}, x_{j_2}, \ldots)$ having the image in $[y_{i_1}, \ldots, y_{i_k}]\cap Y^*$ is in $\mathfrak{A}_X$. We say that $(y_{i_1}, \ldots, y_{i_k})$ is a segment of $(y_{l_1}, \ldots, y_{l_s})$ if $k\leqq s$, and $y_{i_1}=y_{l_1}, \ldots, y_{i_k}=y_{l_k}$. Obviously, the image of $(x_{j_1}, x_{j_2}, \ldots)$ is in $[y_{i_1}, \ldots, y_{i_k}]\cap Y^*$ if and only if $(y_{i_1}, \ldots, y_{i_k})$ is a segment of $c(x_{j_1}, \ldots, x_{j_k})$. Thus

$$d^{-1}([y_{i_1}, \ldots, y_{i_k}]\cap Y^*)=\bigcup[x_{j_1}, \ldots, x_{j_k}], \tag{2}$$

where union runs over sequences $j_1, \ldots, j_k$ for which $(y_{i_1}, \ldots, y_{i_k})$ is a segment of $c(x_{j_1}, \ldots, x_{j_k})$. However the right side of (2) is a union of cylinder sets in $\mathfrak{A}_X$ which proves the Lemma.

Denote by Y^k the set of sequences $(y_{i_1}, \ldots, y_{i_k})$ for which the cylinder set $[y_{i_1}, \ldots, y_{i_k}]$ is in Y^*. Let $\mathfrak{A}_Y^k$ be the σ-field of all subsets of Y^k. By Lemma 1 we can define

$$p_Y^k(y_{i_1}, \ldots, y_{i_k})=p_Y([y_{i_1}, \ldots, y_{i_k}]\cap Y^*).$$

The average information content of the first k signals of the coded sequence is

$$H(\mathscr{Y}^k) = - \sum_{\substack{1 \leqq y_{i_1} \leqq m \\ \vdots \\ 1 \leqq y_{i_k} \leqq m}} p_Y^k(y_{i_1}, \ldots, y_{i_k}) \log p_Y^k(y_{i_1}, \ldots, y_{i_k}),$$

where $\mathscr{Y}^k = (Y^k, \mathfrak{A}_Y^k, p_Y^k)$. Finally the definition of the entropy of $\mathscr{Y}$ is

$$H(\mathscr{Y}) = \lim_{k \to \infty} \frac{H(\mathscr{Y}^k)}{k}$$

(the average information content of one signal of the coded sequence), if this limit exists.

If l_i denotes the length of the sequence $c(x_i)$ $(1 \leqq i \leqq n)$ then the length of $c(x_{i_1}, \ldots, x_{i_k})$ is $\sum_{j=1}^{k} l_{i_j}$. Let L_k be a random variable in the probability space $\mathscr{X}^k$, which takes on the value $\sum_{j=1}^{k} l_{i_j}$ if we have the sequence $(x_{i_1}, \ldots, x_{i_k})$. We say that the average code length is L, if

$$\frac{L_k}{k} \Rightarrow L$$

for $k \to \infty$, where $\Rightarrow$ denotes convergence in probability.

A coding $c(x_i)$ $(1 \leqq i \leqq n)$ is called uniquely decodable, if

$$c(x_{i_1}, \ldots, x_{i_k}) = c(x_{j_1}, \ldots, x_{j_s})$$

holds only in the case $k = s$, $x_{i_1} = x_{j_1}, \ldots, x_{i_k} = x_{j_k}$.

We would like to point out that in the above sequence of definitions only the definitions of entropies, average code length and uniquely decodable coding are important, and the other ones are technical.

Finally, one more definition is necessary to the proof. Denote by Z^N the set of the sequences $(x_{i_1}, \ldots, x_{i_s})$ satisfying the conditions

$$\sum_{j=1}^{s} l_{i_j} \geqq N, \quad \sum_{j=1}^{s-1} l_{i_j} < N$$

Let $\mathfrak{A}_Z^N$ be the σ-field of all subsets of Z^N and put

$$p_Z^N(x_{i_1}, \ldots, x_{i_s}) = p_X[x_{i_1}, \ldots, x_{i_s}].$$

It is easy to see, that

$$\sum_{(x_{i_1}, \ldots, x_{i_s}) \in Z^N} p_Z^N(x_{i_1}, \ldots, x_{i_s}) = 1$$

thus $\mathscr{Z}^N = (Z^N, \mathfrak{A}_Z^N, p_Z^N)$ is a probability space, and

$$H(\mathscr{Z}^N) = - \sum_{(x_{i_1}, \ldots, x_{i_s}) \in Z^N} p_X(x_{i_1}, \ldots, x_{i_s}) \log p_X(x_{i_1}, \ldots, x_{i_s}).$$

Theorems and Proofs

To prove our main theorem we need a lemma.

LEMMA 2. *If L exists,*

$$\left(\frac{H(\mathscr{X}^k)}{L\cdot k}-\frac{H(\mathscr{Z}^N)}{N}\right)\to 0$$

provided that $k=\left[\frac{N}{L}\right]$ *and* $N\to\infty$.

PROOF. Let $H(\mathscr{Z}^N|\mathscr{X}^k)$ be the conditional entropy

$$H(\mathscr{Z}^N|\mathscr{X}^k)=-\sum_{\substack{u\in X^k\\ z\in Z^N}}p_X(u,z)\log p_X(z|u),$$

where $p_X(u,z)$ denotes the probability of the subsets of elements in X^∞, for which the first k elements are $x_{i_1},\ldots,x_{i_k}$ and the first s elements are $x_{j_1},\ldots,x_{j_s}$, if $u=(x_{i_1},\ldots,x_{i_k})$, $z=(x_{j_1},\ldots,x_{j_s})$. Further

$$p_X(z|u)=\frac{p_X(u,z)}{p_X(u)}.$$

Obviously, $p_X(u,z)\neq 0$ $\big(p_X(z|u)\neq 0\big)$ if and only of if one of the u and z is a segment of the other.

It is well known that (e. g. [1])

$$H(\mathscr{X}^k)+H(\mathscr{Z}^N|\mathscr{X}^k)=H(\mathscr{Z}^N)+H(\mathscr{X}^k|\mathscr{Z}^N), \tag{3}$$

so it is sufficient to show, that

$$H(\mathscr{Z}^N|\mathscr{X}^k)=o(N)$$

and

$$H(\mathscr{X}^k|\mathscr{Z}^N)=o(N)$$

if $k=\left[\frac{N}{L}\right]$. Namely, in this case

$$\frac{H(\mathscr{X}^k)}{N}-\frac{H(\mathscr{Z}^N)}{N}=\frac{H(\mathscr{X}^k|\mathscr{Z}^N)}{N}-\frac{H(\mathscr{Z}^N|\mathscr{X}^k)}{N}\to 0$$

follows from (3).

However

$$H(\mathscr{Z}^N|\mathscr{X}^k)=\sum_{u\in X^k}p_X(u)H(\mathscr{Z}^N|u) \tag{4}$$

where $H(\mathscr{Z}^N|u)$ denotes the entropy $-\sum_{z\in Z^N}p_X(z|u)\log p_X(z|u)$. Let $L_k(u)=l(u)$ be the number $\sum_{j=1}^{k}l_{i_j}$ if $u=(x_{i_1},\ldots,x_{i_k})$.

If $l(u)\geqq N$, $p_X(z|u)\neq 0$ can hold only if z is a segment of u (as u cannot be a proper segment of z only if $u=z$), but because of definition of Z^N there can be only one segment of u in Z^N. Thus $p_X(z|u)=1$ for a certain z, and

$$H(\mathscr{Z}^N|u)=0. \tag{5}$$

In the case $l(u)<N$, $p_X(z|u)\neq 0$ if and only if u is a segment of z. On the other hand $s-k\leqq N-l(u)$ since in the case $s-k=N-l(u)$ for the length $l(z)$ the inequality $l(z)\geqq l(u)+s-k=N$ holds. Obviously, we have only $n^{N-l(u)}$ such sequences z, that is,

$$H(\mathscr{Z}^N|u)\leqq \log n^{N-l(u)}=(N-l(u))\log n \tag{6}$$

(as the maximum of the entropy of a distribution on a set of M elements is $\log M$).

Applying (4), (5) and (6) we have

$$H(\mathscr{Z}^N|\mathscr{X}^k)=\sum_{\substack{u\in X^k\\ l(u)<N}} p_X(u)H(\mathscr{Z}^N|u)\leqq \sum_{\substack{u\in X^k\\ l(u)<N}} p_X(u)(N-l(u))\log n=$$

$$=\sum_{\substack{u\in X^k\\ N(1-\varepsilon)\leqq l(u)<N}} p_X(u)(N-l(u))\log n+\sum_{\substack{u\in X^k\\ l(u)<(1-\varepsilon)N}} p_X(u)(N-l(u))\log n\leqq$$

$$\leqq N\varepsilon\log n+p_X\big(u\in X^k,\ l(u)<(1-\varepsilon)N\big)N\log n.$$

Thus we have the inequality

$$\frac{H(\mathscr{Z}^N|\mathscr{X}^k)}{N}\leqq \varepsilon\log n+p_X\big(u\in X^k, l(u)<(1-\varepsilon)N\big)\log n. \tag{7}$$

Since $\dfrac{l(u)}{k}=\dfrac{L_k(u)}{k}$ converges stochastically to L, $p_X\left(u\in X^k, \left|\dfrac{l(u)}{k}-L\right|>\varepsilon\right)$ converges to zero if $k\to\infty$. It follows that on the right side of (7)

$$p_X\big(u\in X^k, l(u)<(1-\varepsilon)N\big)=p_X\left(u\in X^k, \frac{l(u)}{k}-L<(1-\varepsilon)\frac{N}{k}-L\right)\leqq$$

$$\leqq p_X\left(u\in X^k, \frac{l(u)}{k}-L<-L\cdot\frac{\varepsilon}{2}\right)$$

tends to zero for $N\to\infty$, because of $k=\left[\dfrac{N}{L}\right]$. Thus, if N is sufficiently large,

$$p_X(u\in X^k, l(u)<(1-\varepsilon)N)<\varepsilon$$

that is

$$\frac{H(\mathscr{Z}^N|\mathscr{X}^k)}{N}\leqq 2\varepsilon\log n$$

consequently

$$\lim_{N\to\infty}\frac{H(\mathscr{Z}^N|\mathscr{X}^k)}{N}=0\qquad\left(k=\left[\frac{N}{L}\right]\right).$$

We prove in similar way that

$$\lim_{N\to\infty}\frac{H(\mathscr{X}^k|\mathscr{Z}^N)}{N}=0\qquad\left(k=\left[\frac{N}{L}\right]\right).$$

Obviously

$$H(\mathscr{X}^k|\mathscr{Z}^N)=\sum_{z\in Z^N}p_X(z)H(\mathscr{X}^k|z),$$

where $H(\mathscr{X}^k|z)=0$ in the case $z=(x_{i_1}, \dots, x_{i_s})$, $s \geqq k$. Further, if $s<k$, we have only n^{k-s} different u's with $p_X(u|z)\neq 0$, that is, $H(\mathscr{X}^k|z)\leqq (k-s)\log n$. Finally, as above

$$H(\mathscr{X}^k|\mathscr{Z}^N) \leqq \sum_{\substack{z\in Z^N \\ k(1-\varepsilon)\leqq s<k}} p_X(z)(k-s)\log n + \sum_{\substack{z\in Z^N \\ s<(1-\varepsilon)k}} p_X(z)(k-s)\log n \leqq$$
$$\leqq \varepsilon k \log n + k \log n \, p_X\big(z\in Z^N,\, s<(1-\varepsilon)k\big).$$

Here on the right side

$$p_X\big(z\in Z^N,\, s<(1-\varepsilon)k\big) = p_X\big(u\in X^{[(1-\varepsilon)k]},\, l(u)\geqq N\big) =$$
$$= p_X\left(u\in X^M,\, \frac{l(u)}{M}-L \geqq \frac{N}{(1-\varepsilon)k}-L\right) \leqq p_X\left(u\in X^M,\, \frac{l(u)}{M}-L\geqq \varepsilon L\right)$$

which converges to zero if $M=[(1-\varepsilon)k]\to\infty$. Thus $H(\mathscr{X}^k|\mathscr{Z}^N)\leqq o(N)$, indeed, which proves the Lemma.

THEOREM 1. *If the entropy $H(\mathscr{X})$ and the average code length L exist, and the coding is uniquely decodable, then $H(\mathscr{Y})$ exists, and*

$$H(\mathscr{Y}) = \frac{H(\mathscr{X})}{L}.$$

PROOF. If $H(\mathscr{X})$ exists, in Lemma 2 $\dfrac{H(\mathscr{X}^k)}{L\cdot k}$ tends to $\dfrac{H(\mathscr{X})}{L}$, and so $\dfrac{H(\mathscr{Z}^N)}{N}$ does, too. Thus, it is sufficient to show that

$$\lim_{N\to\infty}\frac{H(\mathscr{Z}^N)}{N} = \lim_{N\to\infty}\frac{H(\mathscr{Y}^N)}{N}.$$

We can write

$$H(\mathscr{Y}^N)+H(\mathscr{Z}^N|\mathscr{Y}^N) = H(\mathscr{Z}^N)+H(\mathscr{Y}^N|\mathscr{Z}^N) \tag{8}$$

where

$$H(\mathscr{Z}^N|\mathscr{Y}^N) = -\sum_{\substack{z\in Z^N \\ v\in Y^N}} p_X(v,z)\log p_X(z|v),$$
$$H(\mathscr{Y}^N|\mathscr{Z}^N) = -\sum_{\substack{z\in Z^N \\ v\in Y^N}} p_X(v,z)\log p_X(v|z),$$
$$p_X(z|v) = \frac{p_X(v,z)}{p_Y(v)}, \quad p_X(v|z) = \frac{p_X(v,z)}{p_X(z)},$$

and $p_X(v,z)$ is the probability of the set of sequences in X^∞, for which the first s elements are $x_{i_1}, \dots, x_{i_s}$ and the first N elements of its code are $y_{j_1}, \dots, y_{j_N}$, if $z=(x_{i_1}, \dots, x_{i_s})$, $v=(y_{j_1}, \dots, y_{j_N})$.

Obviously $p_X(v,z)\neq 0$ only if v is a segment of $c(z)$. Thus for given z there is only one v satisfying $p_X(v,z)\neq 0$, that is, $p_X(v,z)=1$. Applying this result we obtain

$$H(\mathscr{Y}^N|\mathscr{Z}^N) = \sum_{z\in Z^N} p_X(z)H(\mathscr{Y}^N|z) = 0 \tag{9}$$

because of $H(\mathscr{Y}^N|z)=0$. On the other hand

$$H(\mathscr{Z}^N|\mathscr{Y}^N) = \sum_{v\in Y^N} p_Y(v)H(\mathscr{Z}^N|v). \tag{10}$$

Let l be the maximum of the numbers $l_1, \ldots, l_n$. Because of definition of Z^N, $N \leqq l(z) < N+l$ holds. Thus for a fixed v, the sequences z satisfying $p_X(z|v) \neq 0$ are such that the first N elements of $c(z)$ are equal to v, and the other elements are arbitrary. The number of such $c(z)$'s is at most m^{l-1}. Since the coding is uniquely decodable i. e. to a given $c(z)$ there exists only one z, the number of different z is also at most m^{l-1}. From (10) we obtain

$$H(\mathscr{Z}^N|\mathscr{Y}^N) \leqq \sum_{v \in Y^N} p_Y(v) \log m^{l-1} = \log m^{l-1},$$

which tends to zero divided by N, if $N \to \infty$. The proof is finished by (8).

If the coding is not necessarily uniquely decodable, we can not prove the existence of $H(\mathscr{Y})$. In this case let us put

$$\overline{H}(\mathscr{Y}) = \overline{\lim_{k \to \infty}} \frac{H(\mathscr{Y}^k)}{k}.$$

In this case from the above proof we get only $\overline{H}(\mathscr{Y}^N) \leqq H(\mathscr{Z}^N)$ that is, the following theorem holds.

THEOREM 2. *If the entropy $H(\mathscr{X})$ and the average code length L exist, then*

$$\overline{H}(\mathscr{Y}) \leqq \frac{H(\mathscr{X})}{L}. \tag{11}$$

Further Questions

1. A natural question is th e following: under which assumption does the limit $\lim_{k \to \infty} \frac{H(\mathscr{Y}^k)}{k}$ exists in the not uniquely decodable case? Probably it is not difficult to answer this question if $\mathscr{X}$ is an information source, which produces independent signals.

2. It is easy to see, that for independent $\mathscr{X}$, and not uniquely decodable coding the strict inequality

$$\overline{H}(\mathscr{Y}) < \frac{H(\mathscr{X})}{L}$$

holds. In other words, in the independent case equality holds in (11) if and only if the coding is uniquely decodable. What is the necessary and sufficient condition, in general, of the equality in (11)?

We are greatly indebted to A. RÉNYI and I. CSISZÁR for several helpful comments and ideas.

REFERENCE

[1] FEINSTEIN, A.: *Foundation of information theory*, McGraw-Hill, New York, 1958.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES,
BUDAPEST

(Received April 1, 1966.)