The principle of conservation of entropy in a noiseless channel.
THE PRINCIPLE OF CONSERVATION OF ENTROPIES
IN A NOISELESS CHANNEL

by

G. Katona and G. Tarsoly

Introduction
The main aim of the paper is to formulate precisely and to prove the following

If we have an information source, name precisely, a sequence of random
variables \( X_1, X_2, \ldots \) with entropy \( H(X) \) and we make its sequence in a perfectly
decidable manner, the obtained sequence \( Y \) has the entropy

\[ H(X) = H(Y) \]

where \( e \) is the average length of the code.

The above equation \( (1) \) of the above represents the principle of conservation of
information, which the coding is uniquely decidable (and no other is possible).
In spite of this, according to our best knowledge \( (1) \) has not been proved in full gener-
ality. If we have the number of code symbols \( s_1, \ldots, s_n \), the maximum of \( H(P) \)
\( s_1, \ldots, s_n \) is the Shannon entropy, which does not depend on the order of
\( s_1, \ldots, s_n \) and

\[ H(P) = \sum_{i=1}^{n} p_i \log_2 \left( \frac{1}{p_i} \right) \]

This is a well-known theorem of Shannon, and according to our best knowledge
with this arrangement of \((1)\) has proved in. \( [24] \).

For the case when the coding is not necessarily uniquely decidable instead of
\( (1) \) we prove the inequality

\[ H(y) \leq H(x) \]

which has an intuitive meaning.

Proof of Theorem
Let \( X = (x_1, \ldots, x_n) \) be the set of possible symbols of the information source, which \( H(X) \) is the sum of all possible probabilities that the state \( x_i \) occurs.

If \( E = \{ x_1, \ldots, x_n \} \) we denote by \( \bar{y} \) the set of all sequences having
\( x_1, \ldots, x_n \), on the first \( n \) positions. We call such subsets of \( E \), called states, \( \gamma \).

The mutual space \( H(x, y) \) is then

\[ H(x, y) = \sum_{i=1}^{n} p_i \log_2 \left( \frac{1}{p_i} \right) \]
The average information content of the first $k$ symbols of the coded sequence is

$$H(P) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

where $p(x)$ is the probability of symbol $x$. Finally, the definition of the entropy of $P$ is

$$H(P) = \sum_{x \in \mathcal{X}} p(x) H(x)$$

The average information content of one signal of the coded sequence is 0 in the limit as $k$ approaches infinity. Let $L$ be the length of the sequence $x_1 \ldots x_k$. Let $L$ be a random variable in the probability space $P$, which takes on the value $\left\lceil \log \frac{1}{p(x)} \right\rceil$ if we have the sequence $x_1 \ldots x_k$. We say that the average code length $L$ is

$$L = -E(-L)$$

where $E(-L)$ denotes the expected value of $-L$. A coding scheme is called uniquely decodable if

$$E(-L) = 0$$

Both $L$ and $E(-L)$ are $\geq 0$. We would like to prove that $L = 0$, i.e., the average number of symbols is 0. A coding scheme is recursive if

$$E(-L) = 0$$

Finally, the average number of symbols is 0. Denote by $E$ the set of sequences $\{x_1 \ldots x_k\}$ satisfying the conditions

$$\sum_{x \in \mathcal{X}} p(x) = 1, \sum_{x \in \mathcal{X}} x = 0$$

Let $W$ be the subset of all symbols of $E$ and put

$$H(w) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

It is easy to see that

$$\sum_{x \in \mathcal{X}} p(x) H(x) = 1$$

and

$$H(P) = \sum_{x \in \mathcal{X}} p(x) H(x)$$

where $p(x)$ is a probability space.
Theorem and Proof.

To prove our main theorem we need a lemma.

Lemma. If \( I \) is open,

\[
\left\{ \begin{array}{c}
\frac{d}{dx} \int_{a}^{b} f(x) \, dx = f(a) \\
\int_{a}^{b} \frac{d}{dx} f(x) \, dx = f(b) - f(a)
\end{array} \right.
\]

provided that \( a < \frac{b}{2} \) and \( \frac{b}{2} < b \).

Proof. Let \( H(F_i) \) be the conditional entropy,

\[
H(F_i) = - \sum_{i} p_i \log p_i.
\]

where \( p_i \) denotes the probability of the subset of elements in \( F_i \) for which the first \( k \) elements are \( x_1, \ldots, x_k \) and the last \( r \) elements are \( y_1, \ldots, y_r \). If \( d_1, \ldots, d_k, r \) and \( y_1, \ldots, y_r \) are independent,

\[
p_i = p(x_1, \ldots, x_k, y_1, \ldots, y_r).
\]

Then, it follows that

\[
\frac{d}{dx} \int_{a}^{b} f(x) \, dx = f(a).
\]

Notice that \( f(x) \) is the probability of the random variable \( x \) taking on a particular value in the interval \( I \).

Proof of the main theorem.

It is well known that (by [10])

(1) \( H(F) = \sum_{i} p_i H(F_i) \),

so it is sufficient to show that

(2) \( H(F) = \sum_{i} p_i H(F_i) \).

If \( i = 1 \),

\[
H(F_1) = - \sum_{i} p_i \log p_i.
\]

Otherwise, \( \sum_{i} p_i H(F_i) \).

Therefore, we have

(3) \( H(F) = \sum_{i} p_i H(F_i) \),

where \( H(F) \) denotes the entropy of \( x \) where \( f(x) \) is defined by \( f(x) = p(x) \). Let \( p_i = 1 \) be the number of elements in the set. For the set \( F_1 \), if \( i \) is a segment of \( x \), then \( x \) is a sequence of \( x \) and \( x \) is a sequence of \( x \). Thus, \( p_1 = 1 \) for the set \( F_1 

\[
\frac{d}{dx} \int_{a}^{b} f(x) \, dx = f(a).
\]
In the case $\eta > 1$, $\eta (\eta - 1) \log(\eta - 1)$ and only $\eta$ is a segment of $\eta$. On the other hand $\eta > 2$. This case in the case $\eta > 2$ and $\eta > 3$. For the length $X$, the distance is $\log(\eta - 1) < -\log(\eta - 1)$, and $\eta > 3$. Therefore, we have only $\eta^{\eta - \log(\eta - 1)}$ and segments $\eta > 3$.

Let $\text{M} = \log(n)$.

We have the maximum of the entropy of a distribution on a set of $N$ elements is $\log(N).$

First, applying the $\log$ and the $\eta$ we have

$$\log(M(t, \eta)) = \log(\log(N)) + \log(M) = \log(N) - \log(\eta) + \log(\log(N)) + \log(M).$$

$$= \frac{\log(M)}{\log(N) - \log(\eta) + \log(\log(N)) + \log(M)} = \frac{\log(M)}{\log(N) - \log(\eta) + \log(\log(N)) + \log(M)}.$$

Thus we have the inequality

$$\log(M(t, \eta)) < \log(N) - \log(\eta) + \log(\log(N)) + \log(M).$$

Now $\frac{\log(M)}{\log(N) - \log(\eta) + \log(\log(N)) + \log(M)}$ converges monotonically to $\log(N) - \log(\eta) + \log(\log(N)) + \log(M).$ It follows that on the right side of (5),

$$\log(M(t, \eta)) < \log(N) - \log(\eta) + \log(\log(N)) + \log(M).$$

This is true for $\eta > 3$, because of $\eta > 3$. Thus, if $\eta$ is sufficiently large, that is

$$\log(M(t, \eta)) < \log(N) - \log(\eta) + \log(\log(N)) + \log(M).$$

Conversely

$$\log(M(t, \eta)) > \log(N) - \log(\eta) + \log(\log(N)) + \log(M).$$

We prove in similar way the

$$\log(M(t, \eta)) > \log(N) - \log(\eta) + \log(\log(N)) + \log(M).$$

Obviously

$$\log(M(t, \eta)) = \log(M) + \log(N).$$
where $W^k = 0$ in the case $k = 1, 2, \ldots, n$. Further, if $\nu = 1$, we have only in the $2^\nu$ columns $e$ with $y^{(2^n)} = 0$, that is, $\nu(y^{(2^n)}) = 0$. Finally,

$$R^\nu(W) = \sum_{i=1}^{2^n} \sum_{j=1}^{2^n} (x_i - x_j)^2$$

is the degree of the second-order polynomial.

But on the right side

$$-2\sum_{i=1}^{2^n} (x_i^2 - \bar{x}^2) = -2\sum_{i=1}^{2^n} x_i^2 + 2nx^2 \bar{x}$$

which reduces to zero if $\bar{x} = 0$. Thus $\nu(y^{(2^n)}) = 0$, indeed, which proves the lemma.

Lemma 1. If the average $\bar{R}(x)$ and the average each length $L$,-code and the coding is uniquely decodeable, then $\bar{R}(x)$ is, and

$$\bar{R}(X) = R(X)$$

Proof. If $R(x)$ is, in Lemma 2,

$$\bar{R}(X) = \frac{\sum_{i=1}^{2^n} R(x_i)}{2^n}$$

does too. Thus, it is sufficient to show that

$$\frac{\sum_{i=1}^{2^n} R(x_i)}{2^n} = \frac{\sum_{i=1}^{2^n} R(x_i)}{2^n}$$

But we now

$$\sum_{i=1}^{2^n} R(x_i) = \sum_{i=1}^{2^n} R(x_i)$$

and the probability of the set of sequences in $W^k$, for which the first $k$ columns are $x_1, \ldots, x_k$, and the last $k$ columns of the code are $x_{k+1}, \ldots, x_{2k}$,

$$P(W^k) = 2^{-kn}$$

is independent of $k$. Thus for given $\alpha$ there is only one possible choice for $k(\alpha)$. Now, applying the method without $\nu(y^{(2^n)})$, we have

$$R(W^k) = \frac{\sum_{i=1}^{2^n} R(x_i)}{2^n} = \frac{\sum_{i=1}^{2^n} R(x_i)}{2^n}$$

because of $\nu(y^{(2^n)}) = 0$. On the other hand

$$\sum_{i=1}^{2^n} R(x_i) = \sum_{i=1}^{2^n} R(x_i)$$

which proves the theorem.
Let $N$ be the maximum of the numbers $1, \ldots, n$. Since each of the numbers $1, \ldots, n$ occurs at least once in $\mathcal{F}$ and the number of these occurrences is at most $n^N$, the total number of outputs of $\mathcal{F}$ is at most $n^N$. The number of all $2^n$-arrays is at most $n^N$. If the coding is uniquely decodable, $F$ is a given (2^n, n)-code only if the number of elements of $F$ is at most $n^N$. From (10) we obtain
\[ H(F) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \leq n^N \log_2 n, \]
which must be an integer. This proves the desired result.

If the coding is not uniquely decodable, we cannot prove the existence of $H(F)$. In this case we put
\[ H(F) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \]

In this case the above proof fails. However, we can still show that
\[ H(F) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \]

In the above proof, we get only $H(F) \leq n^N \log_2 n$, that is, the following theorem holds.

Theorem 2. If the entropy $H(F)$ and the average code length $\ell_c$, then
\[ H(F) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \]

Further Questions:

1. A natural question is the following: under which circumstances does the inequality $H(F) \leq \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$ occur? Probably it is not difficult to answer this question if $F$ is an information source, which produces independent symbols.

2. It is easy to see that for independent symbols and not uniquely decodable coding the exact equality $H(F) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$ 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, for the equality $H(F) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$ to hold? It is a question worth investigating in future research and discussion.

References:
