**Description of the program**

**For any clarification or suggestion to make this explanation clearer, contact us at the following e-mail:** [**Christian.Schmidt55u@gmail.com**](mailto:Christian.Schmidt55u@gmail.com)[**adrain.vdberg66@yahoo.com**](mailto:adrain.vdberg66@yahoo.com)

Definition: Given a sequence x, we call Cs(x) the coded sequence in which the symbols are replaced by a uniquely decodable code.

Codes of this type have the characteristic that no codeword is the prefix of another codeword. Thus, a code is uniquely decodable if it is possible to decode each transmitted character unambiguously, without ambiguity. These codes are very important because many theorems such as Shannon's first theorem refer to this type of codes.

Definition: given a sequence x of random variables of length N, we call the coding limit Lc(x) the function defined as follows:

With we mean the "actual frequency" of the symbol in the sequence.

The function Lc tells us that a sequence x of random variables of length N on average cannot be encoded in less than Lc(x) bits. At most, there may be a function that can transform the sequence x into a new sequence that reduces the value of Lc 50% of the time and increases it the remaining 50%, obtaining no gain on average.

In Set Shaping Theory this function is referred to as the sequence information content. In this article, we have chosen not to use this definition because it can be risky to talk about information without knowing the source that generated the message. This comment does not mean that the definition given in the Set Shaping Theory is incorrect; simply we have chosen a less ambiguous definition linked to an experimental limit.

**The program performs the following experiment:**

Generate a random sequence with uniform distribution (symbol emission probability 1/|A|) with alphabet A and length N.

For example, we generated the following sequence x with N=10 and |A|=5.

2223344551

We calculate the frequencies of the symbols in the sequence.

Symbol 1 frequency 1/10

Symbol 2 frequency 3/10

Symbol 3 frequency 2/10

Symbol 4 frequency 2/10

Symbol 5 frequency 2/10

Calculate the parameter Lc(x) as defined previously.

This value defines the length limit of the encoded sequence in which symbols are substituted for codewords (Uniquely Decodable code).

We perform a transform on sequence x generating a new sequence f(x). Let us take, as an example, a function that transforms the sequence x into the following sequence f(x) with N=11.

11111122435

We calculate the frequencies of the symbols in the sequence.

Symbol 1 frequency 6/11

Symbol 2 frequency 2/11

Symbol 3 frequency 1/11

Symbol 4 frequency 1/11

Symbol 5 frequency 1/11

We encode the sequence f(x) using the Hufmman encoding in which the codewords are optimized for the frequencies that we have calculated (symbol 1 6/11, symbol 2 2/11, symbol 3,4,5 1/11). The codewords found using these frequencies are:

Symbol 1 codeword 1

Symbol 2 codeword 000

Symbol 3 codeword 001

Symbol 4 codeword 010

Symbol 5 codeword 011

The coded sequence becomes:

111111000000010001011

A sequence of 21 bits length, therefore, a value less than the 22.5 bits that constituted the encoding limit Lc(x) of the initial sequence.

The program performs these steps a number of volts indicated by the parameter history. Each time that the transformed sequence f(x) is encoded with a number of bits less than Lc(x) the value of the counter cs increases by one unit. When this cycle ends, we calculate the probability that the transformed sequence f(x) can be encoded Cs(f(x)) with a number of bits less than Lc(x).

The table shows the results for some settings. The first column reports the parameter ns which indicates the number of symbols of the random sequences with uniform distribution generated, therefore represents . The second column reports the length N of the generated messages and the third column the probability ps that the encoding of the transformed message Cs(f(x)) has a length less than Lc(x).

|  |  |  |
| --- | --- | --- |
| ns | N | Pr |
| 40 | 80 | 69% |
| 50 | 100 | 72% |
| 60 | 120 | 80% |
| 500 | 1000 | 88% |

Table 1: Results obtained for different settings of the parameters ns and N.

The results confirm the theoretical predictions, In fact, the transformed sequences f(x) can be encoded with a number of bits lower than the Lc(x) of the initial sequence x with a probability greater than 50%. For example, with |A|=40 and N=80 there is a probability of about 79% that the transformed sequence can be encoded with a bit number lower than the Lc(x) value of the initial sequence x.