**Description of the program**

**Set Shaping Theory**

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

Set Shaping Theory studies the bijection functions that transform a set of strings of length *N* into a set of strings of length *N+K* with K and and .

is the set that contains all the sequences of length N that can be generated from an alphabet A, therefore .

Of particular interest are the functions in which the set contains the sequence with less information content belonging to the set .

The Set Shaping Theory is based on a very simple but extremely interesting experimental result.

Given an alphabet A containing |A| symbols, a number of sequences of length N equal to can be generated, let's call the set that contains all these sequences .

The information content of a sequence of length N and alphabet A is:

is the frequency with which the symbol in position i is present in the sequence.

Now, we calculate the average information content of the sequences belonging to .

For example, if we have and N = 2, the set consists of these four sequences:

11 information content = 0 ()

00 information content = 0 ()

10 information content = 2 ()

01 information content = 2 ()

Hence, the average information content of the sequences belonging to is 4/4 = 1 bit.

The set , that contains the strings with less information content belonging to the set , is the following:

111 information content = 0 ()

000 information content = 0 ()

100 information content = 2.75 ()

011 information content = 2.75 ()

Therefore, the average information content of the sequences belonging to is 5.5/4=1.375 bits.

If we compare the two results, we obtain a result that is not surprising to us; in fact increasing the length of the sequence causes an increase in the average total information content of the sequences. The fact that we have chosen the sequences with less information content does not compensate for their greater length.

Up to now, the results do not surprise us, but if we increase the number of symbols we get an unexpected result, in fact, when |A |>2 the average information content of the sequences at is less than the average information content of the sequences belonging to . The table shows the data in bits for |A| variable from 2 to 10, N=100 and K=1. The data is taken from the following presentation:

<https://www.academia.edu/61997612/Set_Shaping_Theory_the_future_of_information_theory_>

|  |  |  |  |
| --- | --- | --- | --- |
|  | Average inf contet | Average inf contet | difference |
| 2 | 99,275 | 99,659 | -0,383 |
| 3 | 157,044 | 157,040 | 0,004 |
| 4 | 197,819 | 197,324 | 0,495 |
| 5 | 229,271 | 228,304 | 0,968 |
| 6 | 254,843 | 253,401 | 1,443 |
| 7 | 276,353 | 274,464 | 1,889 |
| 8 | 294,868 | 292,527 | 2,341 |
| 9 | 311,121 | 308,383 | 2,738 |
| 10 | 325,570 | 322,388 | 3,181 |

This result is incredible, because the set has the same size as the set . Therefore, through a transform, each element of the set can be placed in one-to-one correspondence with an element of the set .

From this result, I developed the following experiment:

1) generate a random sequence with alphabet A and length N. So, it is as if I randomly extract a sequence from the set .

2) Send the sequence to the encoder that calculate the frequencies of the various symbols in the sequence and compute the total information content of the sequence.

3) Execute the transform f(x)=y with consequently the new sequence is of length N+1.

4) Execute the Huffman encoding and thus obtain the encoded transformed sequence (the symbols replaced by the codewords) and the list of the codewords.

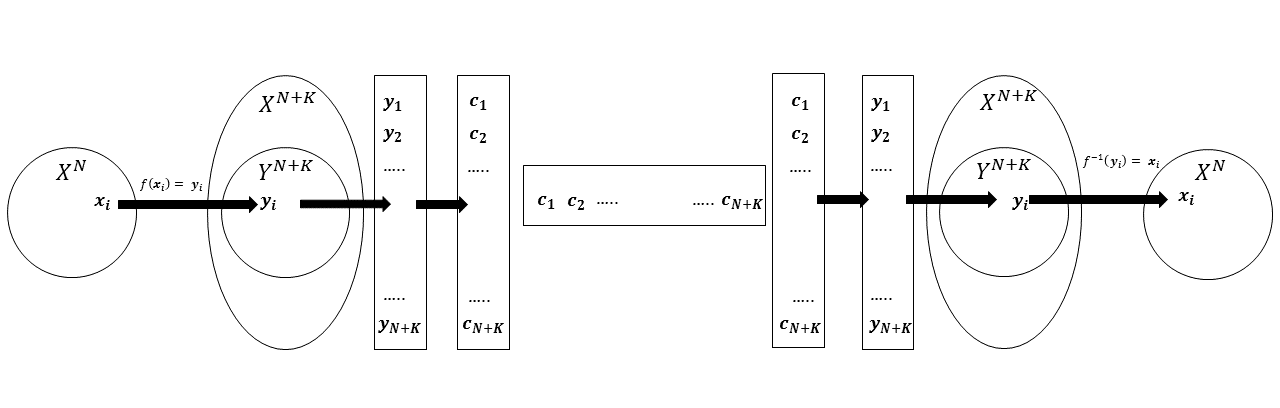
5) Calculate the length of the encoded transformed sequence and compare it with the total information content of the initial sequence x.

6) Sends to the decoder the encoded transformed sequence (the symbols replaced by the codewords) and the list of the codewords.

7) The decoder decodes the sequence and obtains the transformed sequence y.

8) The decoder applies the inverse transform and obtains the initial sequence x.

The figure shows all the steps performed.



The result of this experiment agrees with the data reported in the table. In fact, the transformed sequences can be encoded with a number of bits lower than the information content of the initial sequence 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 information content of the initial sequence.

Furthermore, if, for example, 100 sequences are generated, we have the result that the average encoding length of the transformed sequences is less than the average information content of the generated sequences.

Important: This result by no means that you can compress a random sequence. What is achieved is to reduce the inefficiency of the entropy coding which in the case of the Huffman coding is defined by the list of codewords.