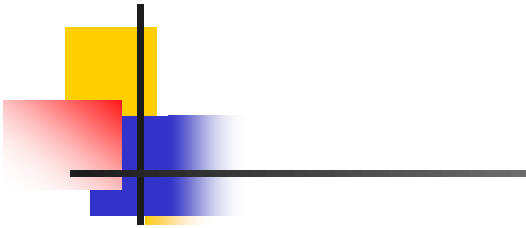
Source Coding & Data Compression



1- Basic Definitions

Source Coding is the representation of source symbols using new alphabet to match the channel alphabet. Coding to binary alphabet is most common source coding as in ASCII code to match binary channel for example. Binary code alphabet {0,1}. Also we may have Ternary code alphabet: {0,1, 2}, Quaternary code alphabet: {0,1, 2,3 }, ….etc.

Source Coding Theorem: The source with entropy H(x) can be represented efficiently with coded alphabet having average length L as long as:

$$L \geq H(x) \qquad (\text{or } L_{min} = H(x))$$

1

with H(x) and L have the same units.

Average Source Coding Length: Suppose we have source symbols X = { x_1 , x_2 , x_3 , ... x_M } with given probabilities: $\{P(x_i)\}$, and the length of each codeword is l_i (number of coded digits to represent x_i). Then the average length of the source coding is given by:

$$\mathbf{L} = \sum_x \boldsymbol{P(x_i)} \, . \, \boldsymbol{l_i}$$

Source Coding Efficiency: The definition of source coding efficiency is:

$$\boldsymbol{\eta_{source\ code}} = \boldsymbol{\eta_{sc}} = \frac{\boldsymbol{H(x)}}{\boldsymbol{L}} \boldsymbol{.100\%}$$

The function of the source encoder is to map each code symbol into corresponding coded word with D-level digits. The mapping should be unique or one-to-one mapping. In the case of binary code (D=2) the source encoder output is connected to 2-level channel or binary channel (as in BSC). The source decoder will map the coded alphabet symbols into source symbols x_i.

2- Main Issues of Source Coding

Consider a source with five symbols as given in the first two columns of the following table. Four different source codes are suggested here. The encoder assigns different codewords to each symbol. Each codeword is represented by one or more alphabet symbols “binary in the example”. The entropy of the source is H(x) =1.875 Bits/Symbol.

Symbol x_i	Prob. $P(x_i)$	Source code#1		Source code#2		Source code#3		Source code#4	
		Codeword	l_i	Codeword	l_i	Codeword	l_i	Codeword	l_i
x_1	0.5	000	3	1010	4	0	1	0	1
x_2	0.25	110	3	11	2	01	2	10	2
x_3	0.125	101	3	0010	4	001	3	110	3
x_4	0.0625	111	3	1	1	1101	4	1110	4
x_5	0.0625	011	3	111	3	0101	4	1111	4
$L = \sum_x P(x_i) . l_i$		3 Bits/Symbol		3.25 Bits/Symbol		1.875 Bits/Symbol		1.875 Bits/Symbol	
$\eta_{sc} = \frac{H(x)}{L}.100\%$		62.5%		57.7%		100%		100%	