UNIVERSITY OF AGRICULTURE, ABEOKUTA

OGUN STATE NIGERIA

Course Code	CSC 417
Course Title	INFORMATION AND COMMUNICATION THEORY
Course Lecturer	Dr. ONASHOGA, S. A. (Mrs.) DEPT. OF COMPUTER SCIENCE UNIVERSITY OF AGRICULTURE, ABEOKUTA, OGUN STATE NIGERIA.

COURSE REQUIREMENTS

This is a compulsory course for all students in the University. In view of this, students are expected to participate in all the course activities and have minimum of 75% attendance to be able to write the final examination.

COURSE CONTENTS

Historical background of information theory, the entropy function and its properties, joint and conditional entropy, discrete memory-less channels, models for communication systems, classification of channels, channel capacity, decoding schemes , the fundamental theorem and its weak converse, finite state channels, continuous channels, entry in the continuous case.

SECTION ONE

FUNDAMENTALS OF COMMUNICATION SYSTEM

In this lesson, student should be able to learn

- the building blocks of a communication system
- the types of communication
- analog communication and digital communication
- transmission impairments
- technology used in communication system.

Discussion: The objective of any telecommunication is to facilitate communication between people who may be closer or located in different corners of the world. The information to exchange can be in different form e.g. text, graphics, voice or videos. In broadcasting information is sent from a central location and people just receive the information. It is not only people, devices may communicate, PC to printer, camera to PC.

1.1 Basis Telecommunication System

The basic principle of all types of communication are the same. The figure below shows a simple telecom system.

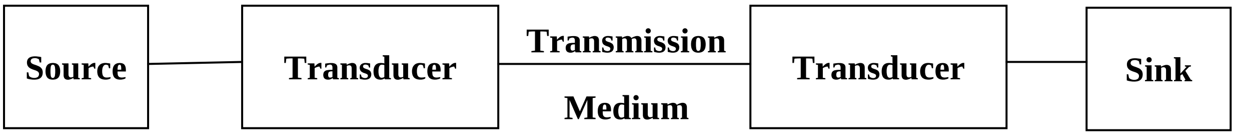


Figure 1: Simple Communication System

In an electrical communication system, at the transmitting side, a transducer converts the real-life information into an electrical signal. At the receiving side, a transducer converts the electrical signal back into real-life information. For example if two people want to talk to each other using this system, the transducer is the microphone that converts the sound waves into equivalent electrical signals. At the receiving end, the speakers convert the electrical signal into acoustic waves. The public address system used in an auditorium is an example of a simple communication system.

The problem with this system is as the electrical signal passes through the transmission medium, the signal gets attenuated. In addition, the transmission medium introduces noise, and as a result, the signal gets distorted.

The noise cannot be eliminated. So the problem is overcome by amplifying the signal and the noise that is added to the actual signal containing the information.

Amplification does not solve the problem particularly when the system has to cover larger distances.

The objective of designing a communication system is for the electrical signal at the transmitting end to be reproduced at the receiving end with minimal distortion. To achieve this, different techniques are used, depending on issues such as type of data, type of communication medium, distance to be covered and so forth.

Different examples of Telecommunication system

(a) **Communication between two computers:** The computer output electrical signals (through serial port, for example) and hence there is no need for a transducer. The data can be passed through the communication medium to the other computer if the distance is small (less than 100 meters.

(b) A communication system in which two PCs communicate with each other over a telephone network. In this system, a new device called a MODEM (Modulator-Demodulator) is introduced at both ends. The PCs send signals, which the modern converts into analog signals and transmits through the medium (copper wires). At the receiving end, the modem converts the incoming analog signal into digital form and passes it on to the PC.

(c) In the case of a radio communication system for broadcasting audio program, the electrical signal is transformed into a high-frequency signal and sent through the air free space. A radio transmitter is used to do this. A reverse of this information – converting the high frequency signal into an audio signal – is performed at the receiving station. Since it is a broadcasting system, many receivers can receive the information.

In a communication system on which two person communicate with two other located somewhere else, but only one communication link, the voice signals need to be combined. The two voice signals cannot be mixed directly because it will not be possible to separate them at the receiving end. There is need to multiplex the two signals, using special techniques.

In a mobile communication system, a radio channel has to be shared by a number of users. Each user has to use the radio channel for a short time during which he has to transmit this data and then wait for his ext turn. This mechanism of sharing the channel is known as multiple access. Hence, depending on the type of communication, the distance to be covered etc. a communication system will consist of a number of elements, each element carrying out a specific function. Some important elements are;

(i) **Muliplexer:** Combines the signals from different sources to transmit on the channel. At the receiving end, a demultiplexer is used to separate the signals.

(ii) **Multiple access:** When two or more users share the same channel, each user has to transmit his signal only at a specified time or using specific frequency band.

(iii) **Error detection and correction:** If the channel is noisy, the received data will have errors. Detection, and if possible correction, of the errors has to be done at the receiving end. This is done through a mechanism called channel coding.

(iv) **Source coding:** If the channel has a lower bandwidth than the input signal bandwidth, the input signal has to be processed to reduce its bandwidth so that it can be accommodated on the channel.

(v) **Switching:** If a large number of users has to be provided with communication facilities, as in a telephone network, the users are to be connected based on the numbers dialed. This is done through a mechanism called switching.

(vi) **Signaling:** In a telephone network, when you dial a particular telephone number, you are telling the network who you want to call. This is called signaling information. The telephone switch (or exchange) will process the signaling information to carry out the necessary operations for connecting to the called party.

1.2 TYPES OF COMMUNICATION

Based on the requirements, the communication can be of different types:

(i) **Point-to-point Communication:** In this type, communication takes place between two end points. For instance, in the case of voice communication using telephones, there is one calling party and one called party. Hence, the communication is point-to-point.

(ii) **Point-to-Multipoint Communication:** In this type of communication, there is one sender and multiple recipients. For example, in voice conferencing, one person will be talking but many others can listen. The message from the sender has to be multilast to many others.

(iii) **Broadcasting:** In a broadcasting system, there is a central location from which information is sent to many recipients, as in the case of audio or video broadcasting. In a broadcasting system, the listeners are passive, and there is no reverse communication path.

(iv) **Simplex Communication:** In simplex communication, communication is possible only in one direction. There is one sender and one sender and one receiver in the sender and receiver cannot change role.

(v) **Half-duplex Communication:** Half-duplex communication is possible in both direction between two entities (computers or person), but one at a time. A walkie-talkie uses this approach. The person who wants to talk presses a talk button on his handset to start talking, and the other person's handset will be in receive mode. When the sender finishes he terminates it with an over message. The other person can press the talk button and start talking. These types of systems require limited channel bandwidth, so they are low cost system.

(vi) **Full-duplex communication:** In a full-duplex communication system, the two parties – the caller and the called – can communicate simultaneous as in a telephone system. However, note that the communication system allows simultaneous transmission of data, but when two person talk simultaneously there is no effective communication. The ability of the communication system to transport data in both directions defines the system as full duplex.

Depending on the type of information transmitted we have voice communication, data communication, fax communication, and video communication system. When various type of information are clubbed together we talk of multimedia communication.

With the advent of digital communication and "convergence technologies", this distinction is slowly disappearing and multimedia is becoming the order of the day.

1.3 TRANSMISSION IMPAIRMENTS

While the electrical signal is traversing over the medium, the signal will be impaired due to various factors. These transmission impairments can be classified into 3 types:

a) Attenuation distortion

b) Delay distortion

c) Noise

The amplitude of the signal wave decreases as the signal travels through the medium. This effect is known as attenuation distortion.

Delay distortion occurs as a result of different frequency components arriving at different time in the guided media such as result copper wire or coaxial cable.

Noise can be divided into 4 but to discuss 2

Cross talk: Unwanted coupling between signal paths is known as cross talk. In the telephone network, this coupling is quite common. As a result of this, other conversation are heard. Cross talk is eliminated by using approximate design techniques.

Impulse noise: This is caused by external electromagnetic disturbance such as lighting. This noise is unpredictable. When the signal is traversing the medium, impulse noise may cause sudden bursts of errors. This may cause a temporary disturbance in voice communication. For data communication, appropriate methods need to be devised whereby the lost data is retransmitted.

1.4 ANALOG VERSUS DIGITAL TRANSMISSION

The electrical signal output from a transducer such as microphone or a video camera is a analog signal, that is, the amplitude of the signal varies continuously with time. Transmitting this signal (with necessary transformations) to the receiving end results in analog transmission.

The output of a computer is a digital signal. The digital signal has a fixed number of amplitude levels. For instance, binary I can be represented by one voltage level (say 5 volt) and binary 0 can be represented by analog level (say 0 volt). The voice and video signals (output of the transducer) are always analog.

Analog signal is converted to digital format through analog digital conversion.

Digital transmission is much more advantageous than analog transmission because digital systems are comparatively immune to noise. The advantages of digital systems are.

i) More reliable transmission because only discrimination between ones and zeros is required.

ii) Less costly implementation because of the advances in digital logic chips.

iii) Ease of combining various of signals (voice, video etc)

iv) Ease of developing secure communication system. All the newly developed communication systems are digital system. Only in broadcasting application is analog communication used extensively.

Assignment

Write a report on the history of telecommunication, listing the important milestones in the development of telecommunication technology.

Class Test

Differentiate between simplex communication and full-duplex communication

SECTION TWO

INFORMATION THEORY

In this lesson, student would learn;

- The Requirements of a communication system.
- The Building blocks of a communication as proposed by Shannon.
- Entropy and channel capacity
- Shannon's source coding Theorem.

2.1 REQUIREMENTS OF A COMMUNICATION SYSTEM

The requirement of a communication system is to transmit the information from the source to the destination without errors, in spite of the fact that noise is always introduced in the communication medium.

2.1.1 The Communication System

Figure 2 shows a generic communication system.

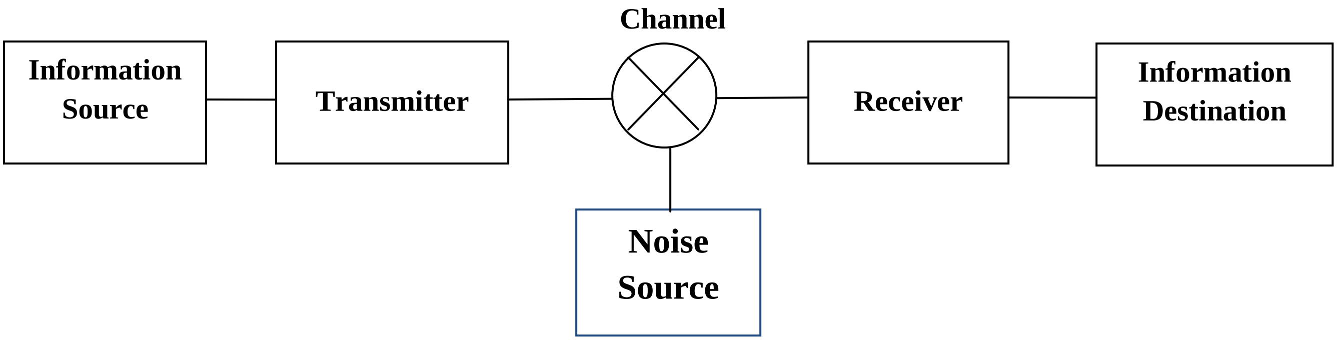


Figure 2: Generic Communication system

The information source produces symbols (such as English letters, speech, video etc) that are sent through the transmission medium by the transmitter. The communication medium introduces noise, and so errors are introduced in the transmitted data. At the receiving end, the receiver decodes the data and gives it to the information destination.

Example 2.1

As an example, consider the information source that produces two symbols A and B. The transmitter codes the data into a bit stream. For examples, A can be coded as 1 and B as 0. The stream of 1s and 0s is transmitted through the medium. Because of noise, 1 may become 0 and 0 may become 1 at random places as illustrated below:

Symbols produced:	A	B	B	A	A	A	B	A	B	A
Bit stream produced:	1	0	0	1	1	1	0	1	0	1
Bit stream received:	1	0	0	1	1	1	1	1	0	1

At the receiver, one bit is received in error. How to ensure that the received data can be made error free is now provided by Shannon.

Hence the generic communication system is now expanded as shown overleaf:

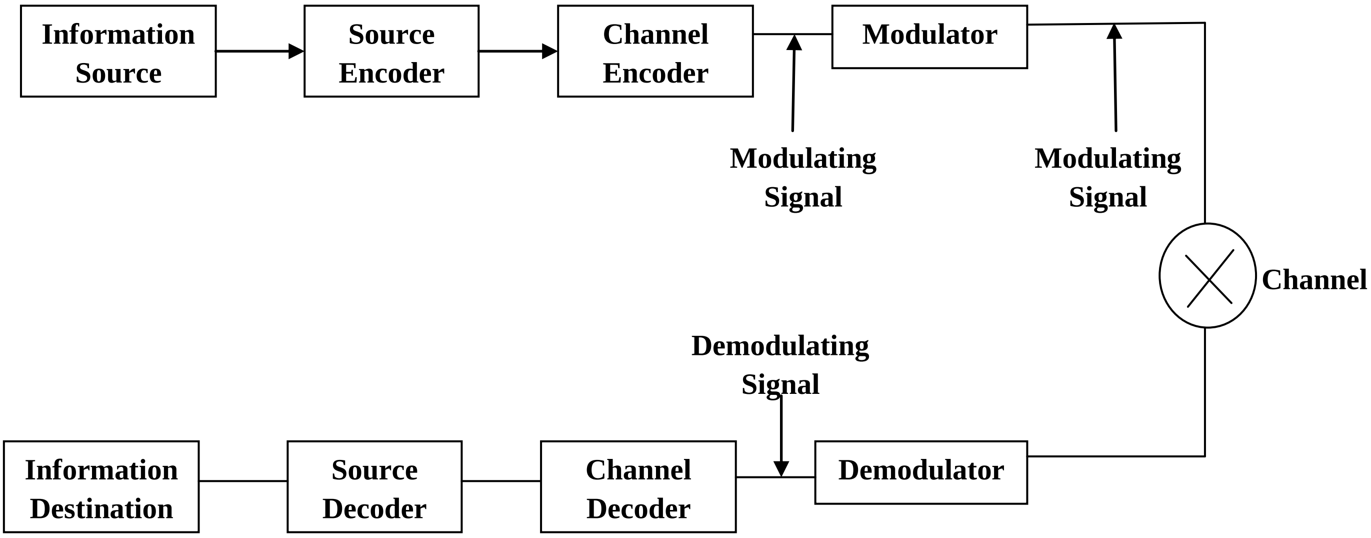


Figure 3: Generic communication system as proposed by Shannon.

In figure 3, the information source produces the symbols that are coded using two types of coding: source encoding and channel encoding and then modulated and sent over the medium. At the receiving end, the modulated signal is demodulated, and the inverse operations of channel encoding and source encoding (channel decoding and source decoding) are performed. Then the information is presented to the information sink. Each block is explained as

I. **Information Source:** The information source produces the symbols. If the information source is for example, a microphone, the signal is in analog is in digital form (a set of symbols).

The theory

Let S be the source and P be the probability associated with it; that is for q source

$S_1, S_2, \ldots S_q$: Source alphabet

$P_1, P_2, \ldots P_q$: Probability

Facts:

(1) The information content $I(S_i)$ is inversely proportional to the probability of occurrence, Pi.

$$\text{i.e.} \quad I(S_i) \propto \frac{1}{Pi} \text{\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} (1)$$

(2) Recall from the law of probability distribution, if 2 random variables X and Y are stochastically dependent then

$$P(XY) = (P(X) P(Y)$$

Thus, the information content from two different independent symbols is the sum of the information content from each separately i.e.

from (1)

$$I(Si) = \log\left(\frac{1}{Pi}\right) = -\log Pi \text{\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_} (2)$$

For 2 different independent symbols; we have

$$I(S_1, S_2) = I(S_1) + I(S_2)$$

$$= \log\left(\frac{1}{P_1}\right) + \log(\frac{1}{P_2}) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3)$$

Since $\log_b a + \log_b C = \log_b aC$

then (3) becomes

$$\log\left(\frac{1}{P_1 P_2}\right) = I\ (S_1\ S_2)$$

II. Source Encoder

The source encoder converts the signal produced by the information source into a data stream. If the input signal is analog, it can be converted into digital form using analog-to-digital converted. If the input to the source encoder is a stream of symbols, it can be converted into a stream of 1's and 0s using some type of coding mechanisms.

For instance if the source produces the symbols A and B, A can be coded as 1 and B as 0.

Source encoding is done to reduce the redundancy in the signal. The two types of source coding techniques are lossless coding and lossy coding. In lossless coding technique, there is no loss of information e.g. when a computer file is compressed using a compression technique for instance WinZip. Compression is the process of coding to effectively reduce the total number of bits needed to represent certain information. It is the process of making data representation smaller, in order to decrease the data's bandwidth and storage requirement. Compression is a major contribution to data communication. It reduces the cost and resources necessary for transmission by reducing both the transmission time and network payload for the same amount of information. It is achieved by redundancy; extra information that is necessary is removed.

In lossy coding technique, some information is lost while doing the source coding. As long as the loss is not significant, the data can still be tolerated. When an image is converted into JPEG format, the coding is lossy coding because some information is lost. Most of the technique used for voice image and video coding are lossy coding techniques.

Take home:

1. List out the techniques/algorithm for both lossy and lossless compression.
2. Write a program to generate a bit stream of 1s and 0s.

III. Channel encoding

To decode the information correctly, even if errors are introduced in the medium, there is need to put some additional bits in the source-encoded data so that the additional information can be used to detect and correct the errors. This process of adding bits is done by channel encoder.

In channel encoding, redundancy is introduced so that the receiving end, the redundant bits can be used for error detection or error correction.

IV. **Modulation:** This is the process of transforming the signal so that the signal can be transmitted through the medium.

V. **Demodulation:** The demodulator performs the inverse operation of the modulator

VI. Channel decoder

The channel decoder analyzes the received bit stream and detects and corrects the errors if any, using the additional data introduced by the channel encoder.

VII. Source decoder

The source decoder converts the bit stream into the actual information. If the analog-to-digital conversion is done at the source encoder, digital-to-analog conversion is done at the source decoder.

If the symbols are coded into 1s and 0s at the source encoder, the bit stream is converted back to the symbols by the source decoder.

VIII. **Information Destination/Sink:** The information sink absorbs the information.

2.2 ENTROPY OF AN INFORMATION SOURCE

In information theory, an information source is a probability distribution, i.e., a set of probabilities assigned to a set of outcomes. This reflects the fact that the information contained in an outcome is determined not only by the outcome but by low uncertain it is. An almost certain outcome contains little information.

Example

Consider a rain forest in the Sahara desert. This forecast is an information source. The information source has two outcomes; rain or no-rain. Clearly, the outcome no-rain contains little information; it is a highly probable outcome. The outcome ran, however, contain considerable information, it is a highly improbable event.

Shannon proposed a formula to measure information. This information measure is called the entropy of the source. It is a measure of the amount of uncertainty in a probabilities choice system (S, P) (the choice, S together with its sets of probabilities).

If a source produces N symbols and if all the symbols are equally likely to occur, the entropy of the source is given by

$$H = \log_2 N \text{ bits/symbol}$$

For example, assume that a source produces the English letters (A to Z and space, totally 27, as symbols), and all these symbols will be produced with equal probability. In such a case, the entropy is

$$\begin{aligned} H &= \log_2 27 \\ &= 4.75 \text{ bits/symbol} \end{aligned}$$

The information source may not produce all the symbols with equal probability. For instance, in English the letter “E” has the highest frequency (and hence highest probability of occurrence), and the other letters occur with different probabilities. In general if a source produces i^{th} symbol with or probability of P(i), the entropy of the source is given by

$$H (S) = -\sum P(i) \log_r P(i)$$

It can be defined as the average information content over the whole alphabet of symbols

that is $Hr (S) = \sum Pi\ log_r\ (\frac{1}{P_i})$ __________________ (4)

$$= \sum Pi\ log_r\ (P_i)^{-1}$$

$$= -\sum_{i=1}^{q} P_i log_r\ (P_i)^{-1}$$

This can be represented as

$Hr (S) = H_2 (S)\ log_r 2$ _______________________ (5)

To show that (4) and (5) are equal

From (5) $H_2 (S)\ log_r 2$

$\sum P_i\ log_2 \left(\frac{1}{Pi}\right)\ log_r 2$ _______________________(6)

Recall $\log_b a = \frac{log_c a}{log_c b}$

from (6) we have $log_2 (\frac{1}{Pi}) = \frac{log_r\ (\frac{1}{Pi})}{log_r (2)}$

⇒ (6) becomes

$$\sum \frac{Pi\ log_r\ (\frac{1}{Pi})}{log_r 2}\ log_r 2$$

$= \sum_{i=1}^{q} Pi\ log_r\ (\frac{1}{Pi})$ ________________ Proved

Entropy is a function of the probability distribution Pi and does not involve Si. Its unit of measurement is bits/symbols, hence the base, r = 2.

Class work

1) A source produces two symbols A and B with probabilities of 0.6 and 0.4 respectively. Calculate the entropy of the source.

$$H = -\sum Pi \log Pi$$

$$= -(0.6\ log_2 0.6 + 0.4\ log_2\ 0.4)$$

= 0.970 bits/ symbol

2) A source produces 42 symbols with equal probability. Calculate the entropy of the source.

$$\text{H} = log_2 42 \text{ bits/symbol}$$

$$= 5.55 \text{ bits/symbol}$$

3) Consider a source that produces four symbols with probabilities of $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ and $\frac{1}{8}$ and all symbols are independent of each other. Calculate the entropy.

Solution

$$\text{H} = - \sum P_i \log P_i$$

$$= \{\frac{1}{2}\ log_2 \frac{1}{2} + \frac{1}{4}\ log_2\ \frac{1}{4} + \frac{1}{8}\ log_2\ \frac{1}{8} + \frac{1}{8}\ \ log_2 \frac{1}{8} \quad \text{OR}$$

$$\text{H} = \sum Pi \log \frac{1}{Pi}$$

$$= \frac{1}{2}\ log_2 2 + \frac{1}{4}\ log_2 4\ + \frac{1}{8}\ log_2 8\ + \frac{1}{8}\ \ log_2 8$$

$$= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8}$$

$$1\ + \frac{3}{4} = \frac{7}{4} \text{ bits/symbol.}$$

4) Prove that for a certain event H(X) = 0

2.2.1 Properties of Entropy

Recall from Probability Distribution Theory:

Let X and Y be 2 independent random variables

then $P\ (XY) \neq P(X)P(Y)$ _____________ Joint Distribution

$P\ (X) = \ \sum_{y_{iEY}} P\ (X, Y = yi) \Rightarrow$ Marginal distribution

$P\ (X/Y) = \ \sum_{y_{iEY}} P\ (X, Y = yi) P\ (X, Y = yi) \Rightarrow$ Conditional distribution

By Baye's rule

$$P\ (X, Y) = \ P(X/Y)\ P(Y)$$

$$= \ P(Y/X)P(X)$$

(1) If (S_1, P_1) and (S_2, P_2) are two probabilistic choice with $|S_1| = |S_2|$ and $P_1 = P_2$, then $H(P_1) = H(P_2)$. This implies that entropy depends only on the probability distribution but not on the nature of the possibilities in the choice system.

(2) $H (P_1, P_2 \ldots, P_n) = H (P_1, P_2, \ldots P_n, 0)$

From the convention 0.log0 = 0. It says that possibilities with various probabilities are irrelevant to the amount of uncertainty.

(3) $H(X) \geq 0$ with equality iff Pi = 1

(4) Joint entropy

$$H(X,Y) = \sum_{x \in X} \sum_{y \in Y} P\ (x,y) \log \frac{1}{P(x,y)}$$

(5) Conditional Entropy

$$H(X/Y)) = -\sum_{y_i \in Y} P\ (Y = yi)\ H(X/Y = yi)$$

Example: to show further the properties

For a set of possible message $X = (x_{1,}x_{2,} \ldots\ x_N)$ with equal probability $\{P\!: Pi = P$ for i = 1, $N\}$

i.e. $P = \frac{1}{N}$ $\{e.g\ Dice = \frac{1}{6}\}$

$$H(X) = -\sum P \log P$$

$$= -\sum \frac{1}{N} \log \frac{1}{N}$$

$$\Rightarrow H\left(x_i, x_j\right) = -\sum_{i,j=1} \frac{1}{N^2}\ log_2\ \frac{1}{N^2}$$

$$= -\sum \frac{1}{N^2}\ log_2\, N^{-2}$$

$$= \sum \frac{1}{N^2}\ log_2\, N^2$$

$$= log_2 N^2$$

= 2 $log_2 N$

= $log_2 N + log_2 N$

i.e. for two symbols with equal probabilities

H (x_i, x_j) = $\log_2$ N + $\log_2$N

Take Home: Show that H(X,Y) ≤ H(X) + H(Y)

On Property (5) - Conditional Entropy

Consider two choice system S_1 and S_2 and the associated system of independent choices.

$S_1 x\, S_2 = \{(e_{1,1}, e_{2,1})(e_{12} e_{22}) \ldots (e_{1n} e_{2n})\}.$

By affecting probabilities $P_{i,j}$ to the compound choice $(e_{1i,} e_{2j})$ this system of independent choices can be extended to a compound probabilistic choice system (S_1 x S_2, P) where P = {$P_{i,\,j}$: i = 1 … n, j = 1 … m}

$\Rightarrow \sum_{i=1}^{m} \sum_{j=1}^{m} Pi, j = 1$ 0 ≤ Pi, j

Let x be associated with the system (S_1, P_1) and Y with the system (S_2, P_2). The pair of variables (X, Y) is then associated with the compound probabilistic system (S_1 x S_2, P).

Recall Two r.v X and Y are independent iff

$P_{x,y}$ (x, y) = P(x). P(y)

There are 3 entropies associated with the 3 probabilistic choice system X, Y and (X, Y)

i.e. $H\,(X, Y) = -\sum_{xES_1} \sum_{yCS_2} P_{X,Y}\,(x, y) \log P_{X,Y}\,(x, y)$

$H\,(X) = -\sum_{xES_1} P_x\,(x) \log P_x\,(x)$

$H\,(Y) = -\sum_{yES_2} P(y) \log P(y)$

This theorem can be used to show that

H (X, Y) ≤ H (X) + H (Y)

H (X) + H (Y)

$= -\{\sum P\,(x) \log P(x) + \sum P(y) \log P(y)\}$

$= -\{\sum\sum P(x,y)\log(P(x) + \sum\sum P(x,y)\log P(y)\}$

$= -\{\sum_x \sum_y P(x,y)\log P(x).P(y)\}$

From this lemma $\sum P(i)\log q(i) \leq \sum P(i)\log P(i)$

$$H(X) + H(Y) \geq -(\sum\sum P(x,y)\log P(x,y)$$

i.e. $-\{\sum_x \sum_y P(x,y)\log P(x).P(y)\} \geq -\{\sum\sum P(x,y)\log P(x,y)\}$

i.e. $H(X,Y) \leq H(X) + H(Y)$

Example: Given a compound system of independent choices

S_1 x S_2 = {e_{11}, $e_{2,1}$) ($e_{11}e_{22}$) ($e_{12}e_{21}$), ($e_{12}e_{22}$)}

P = {0.5, 0.1, 0.3, 0.1}

and two random variables (X, Y) associated with (S_1, S_2, P) with the single choice system identified as S_1 = {e_{11}, e_{12}}, S2 = {e_{21}, e_{22}}

with P_1 = {0.6, 0.4} and P_2 = {0.8, 0.2}

the entropies are now calculated as

$H(X,Y) = -\sum\sum P(x,y)\log P(x,y)$

$= -0.5\log 0.5 - 0.1\log 0.1 - 0.3\log 0.3 - 0.1\log 0.1$

$\simeq$ 1.6855 bit

$H(X) = -\sum P(x)\log P(x)$

$= -0.6\log 0.6 - 0.4\log 0.4$

$\simeq$ 0.9710 bit

$H(X) = -\sum P(y)\log P(y)$

$= -0.8\log 0.8 - 0.2\log 0.2$

$\simeq$ 0.7219 bit

which clearly shows the proof $H(X,Y) \leq H(x) + H(Y)$

For two variables X and Y. Suppose the value of one variable say Y = y is observed. How does this affect the uncertainly concerning variable X?

This observation changes the distribution P(x) to the conditional distribution $P_{x/y}$ *(x,y)* defined as

$$P_{x/y}(x,y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *$$

Therefore, the conditional entropy of X given Y-y is given as

$$H\ (X/Y = y) = -\sum_x P_{\frac{x}{y}}(x,y)\log P_{x/y}\ (x,y)$$

i.e. $H\ (X/y) = -\sum P_{x/y}\ (x,y)\log P_{x/y}\ (x,y)$ ______________**

Note however, that if the two variables are independent then we have $P_{\frac{x}{y}}(x,y) = P_x(x)$ for every x and y.

In this case

$$H\ (X/Y) = -\sum_x P_x\ (x)\log P_x(x) = H\ (x)$$

The uncertainty in X does not change, when a variable Y which is independent of X is observed.

Example 2 on Conditional Entropy

Given $P_{X,Y}$ (0,1) = $P_{X,Y}$ (1,0) = $P_{X,Y}$ (0,0) = $^1/_3$

$P_{X,Y}$ (1,1) = 0, P_x (0) = P_Y (0) = $^2/_3$

P_X (1) = P_Y (1) = $^1/_3$.

Hence H (X/Y) = $-\sum P_{x/y}\ (x,y)$ log $P_{x/y}$ (x,y)

$H\ \left(\frac{X}{Y} = 0\right) = -P_{x/y}$ (0,0) log $P_{x/y}$ (0,0) – $P_{x/y}$ (1,0) log $P_{x/y}$ (1,0)

$$= \frac{P_{X,Y}\ (0,0)}{P_Y\ (0)} \log \frac{P_{X,Y}\ (0,0)}{P_Y\ (0)} - \frac{P_{X,Y}\ (1,0)}{P_Y\ (0)} \log \frac{P_{X,Y}\ (1,0)}{P_Y\ (0)}$$

$$= -\frac{1/3}{2/3} \log 1/3 \div 2/3 - 1/3 \div 2/3 \log 1/3 \div 2/3$$

$$= -5.5 \log 0.5 - 0.5 \log 0.5$$

$$= 1$$

$$H\ (X/Y = 1) = \frac{P_{X,Y}\ (0,1)}{P_Y\ (1)} \log \frac{P_{X,Y}\ (0,1)}{P_Y\ (1)} - \frac{P_{X,Y}\ (1,1)}{P_Y\ (1)} \log \frac{P_{X,Y}\ (1,1)}{P_Y\ (1)}$$

$$= 1/3 \div 0 \ldots - 0$$

$$= 0$$

$$H\ (X) = H\ (Y) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$$

$$= 0.9183$$

$\therefore \quad H\ (X|Y = 1) < H\ (X) < H\ (X \mid Y = 0)$

Property 5b: Expected Conditional Entropy

In addition to the conditional entropy of X given a particular observation Y = y, the expected conditional entropy of X given Y, which is the expected value of H (x | y) relative to y can be considered as

$$H\ (X|Y) = \sum_y P_Y(y) H\ (X \mid y)$$

Note that

$P_{XY}\ (x,y) = P_{(y)}\ P_{x|y}\ (x,y)$

Therefore, the expected conditional entropy is as follows:

$H\ (X|Y) = \sum_y P_Y(y) H\ (X \mid y)$ ___________ to expand $H_{(x|y)}$

From **

$$H\ (X|Y) = \sum_{x}\sum_{y} P_Y\ (y)\ P_{x\mid y)}\ (x,y) \log P_{x\mid y}\ \ (x,y)$$

$$= -\sum_{x}\sum_{y} P_{X,Y}\ \ (x,y)(x,y) \log \frac{P_{X,Y}(x,y)}{P_Y(y)} \text{from} *$$

$$= -\sum_{x}\sum_{y} P_{X,Y}\ \ (x,y)\ (\log P_{X,Y}\ (x,y) - \log P_Y\ \ (y)$$

$\Rightarrow \qquad H\ (X|Y) = H\ (X,Y) -\ H\ (Y)$

Thus, gives the theorem for a pair of r.v X and Y

$H\ (X,Y) = H\ (Y) +\ H\ (X/Y)$

This theorem says the uncertainty of a pair of variables can be considered as the result of a chaining, where the uncertainty of one of the valiables is considered first, say Y and then add the (expected) conditional uncertainty of the second one given the first one.

It can start with any of the variable. Thus

$$\text{H (X, Y) = H (X) + H (Y} \mid \text{X)}$$

This says that the whole uncertainty in the system is composed of the uncertainty of the input signal H (X) and the transmission uncertainty over the channel H (Y/X)

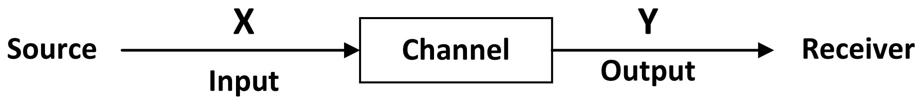


Figure 4: Transmission channel

Example 2.2.1(b) On Expected Conditional Entropy

From example 2.2.1(a)

$$H\ (X|Y) = P_Y(0)H\ (X|Y = 0) +\ P_Y\ (1)H\ (X\mid Y = 1)$$

$$= {}^{2}/_{3}\ (1) +\ {}^{1}/_{3}\ (0)$$

$$= \mathbf{{}^{2}/_{3}\ bit}$$

$H\ (X,Y) = \ -\log {}^{1}\!/_{3} \simeq 1.5850\ $ bit

Check : $H\ (X,Y) = H\ (Y) + \ H\ (X \mid Y)$

	$H\ (Y) =$	0.9183
+	$H\ (X\vert\ Y) =$	0.6667
	$H\ (X,\ Y)$	**1.5850**

$$H\ (X,Y) = \ -\sum\sum P(x,y)\log P(x,y)$$

$= -[P(0,0)\log P(0,0) + \ P(1,0)\log P(1,0) + \ P(0,1)\log P(0,1) + \ P(1,1)\log P(1,1)]$

$= -[\frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3} + \ 0]$

$= -3\ [{}^{1}\!/_{3}\ log_{2}\ {}^{1}\!/_{3}]$

$= \ -log_{2}\ {}^{1}\!/_{3} = \ log_{2}3 = \frac{\log\ 10^{3}}{\log\ 10^{2}} = \simeq 1.5850\ bit$

2.2.2 Mutual Information

Consider a communication channel, with X the input source and Y the output. By observing the output Y we expect a positive amount of information regarding the input X. Although in particular transmission, uncertainty about the input is increased, on average the observation of the output decreases uncertainty about the input.

Given a compound probabilistic choice system (S_1 x S_2, P) and the associated random variables X and Y, the mutual information between these 2 variables can be defined as the expected gain in information on one variable, obtained if the other is observed. It is also the difference between the sum of the individual entropies of the 2 variables and the actual entropy of the pair. This implies that mutual information is always non-negative i.e.

I (X, Y) $\geq$ 0 iff the 2 variables X, Y are independent.

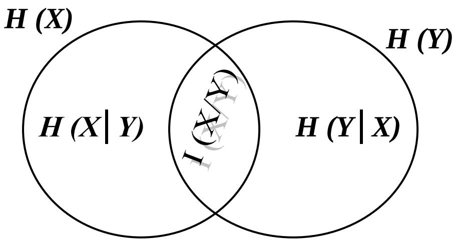
i.e. I (X, Y) = H (P(X) – H (P(X $\vert$ Y)

OR $\qquad$ I (X, Y) = H (X) – H (X │ Y)

The goal is to formally quantity the reduction in uncertainty by examining the appropriate subtraction of entropies.

Illustration of Mutual Information

Using the Venn diagram



From this Venn diagram, mutual information can be defined *I (X│ Y)* or *I (X, Y)*

I (X│ Y) = H (X) – H (X│ Y)

I (X│ Y) = H (Y) – H (Y│X)

I (X│ Y) = H (X) + H (Y) – H (X, Y)

Because this formula is symmetric in X and Y, then *I (X│ Y) = I (Y│X)*

So, mutual information between X and Y is the same as between Y and X.

Example 2.2.2

Given the following probability distribution for transmitting information over a channel:

P(X) = $\{\frac{1}{2}, \frac{1}{2}, 0\}$

$$P\ (X|Y) = 0) = \{1 - P, P, 0\}$$

$$P\ (X|Y) = 1) = \{P, 1 - P, 0\}$$

Let the variable, P = 0.2 be the probability of incorrect transmission. Calculate the mutual information on this distribution.

Solution: I (X, Y) = H (X) – H (X│Y)

Given are Probability of incorrect transmission i.e. Y = 1 and of correct transmission, Y = 0

Calculate all the entropies:

$$\text{H (X)} = -\sum Pi \log Pi$$

$$H\ (X) = -(\tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{2}\log\tfrac{1}{2} + 0)$$

$$= -\left[-\tfrac{1}{2} - \tfrac{1}{2}\right] = 1\ \text{bit/symbol}$$

$$H\ (X|Y) = y) = -\sum P\ (Y = yi) H\ (X|Y) = yi)$$

From ** $H\ (X|Y = 1) - [P\ log_2 P + (1-P) log_2\ (1-P) + 0]$

$$= -[(1-P) log_2\ (1-P) + P\ log_2\ P]$$

$$= (1-P) log_2\ (1-P)^{-1} + P\ log_2\ P^{-1}]$$

$$= (1-P) log_2 \left(\tfrac{1}{1-P}\right) + P\ log_2\ ({}^{1}/_{P}) \qquad \_\_\_\ \text{(i)}$$

$$H\ (X|Y = 0) = -[(1- P) log_2 (1-P) + P\ log_2\ P + 0]$$

$$= (1-P) log_2 \left(\tfrac{1}{1-P}\right) + P\ log_2\ ({}^{1}/_{P})$$

$$\text{I}\ (XiY) = H\ (X) - [P\ (Y = 0) H\ (X|Y\ 0) + P(Y = 1) H(X|Y = 1)]$$

$$= 1 - [{}^{1}/_{2}\ [(1-P) log_2 \left(\tfrac{1}{1-P}\right) + P\ log_2 ({}^{1}/_{P})]]$$

$$+ {}^{1}/_{2} (\ (1-P) log_2 \left(\tfrac{1}{1-P}\right) + P\ log_2 ({}^{1}/_{P})\ )$$

$$= 1 - [{}^{1}/_{2}\ [(1-P) log_2 \left(\tfrac{1}{1-P}\right) + {}^{1}/_{2}\ P\ log_2 ({}^{1}/_{P}) + {}^{1}/_{2} \left[(1-P) log_2 \left(\tfrac{1}{1-P}\right) + {}^{1}/_{2}\ P\ log_2 + ({}^{1}/_{P})\right]$$

$$= 1 - \left[(1-P) log_2 \left(\tfrac{1}{1-P}\right) + P\ log_2 ({}^{1}/_{P})\right]$$

where P = 0.2

$$= 1 - \left[0.8\ log_2 \left(\frac{1}{0.8}\right) + 0.2\ log_2 ({}^{1}/_{0.2})\right]$$

$$= 1 - [0.8\ log_2\ 1.25 + 0.2\ log_2\ 5]$$

$$= 1 - \left[0.8\ \frac{log_{10} 1.25}{log_{10} 2} + 0.2\ \frac{log_{10} 5}{log_{10} 2}\right]$$

$$= 1 - \left[0.8\ \frac{0.0969}{0.30103} + \quad 0.2\ \frac{0.69897}{0.30103}\right]$$

$$= 1 - 0.8\ (0.321895) + \ 0.2\ (2.32193$$

$$= 1 - (0.257516 + 0.4643856)$$

$$= 1 - 0.721902$$

$$= 0.278098$$

2.3.3 Kullback-Leibler Divergence

There is another important notion of information theory. Considered two probabilistic choice system, with the same choice set S, but different probability distributions. If X and Y are the corresponding r.v., this definition

$K = (P_{X,} P_Y = \ \sum_{xES} P_X(x) \log \frac{P_X(x)}{P_Y(x)}$ ________________ ***

This is called the Kullback-Leibler divergence between X and Y. It is a kind of distance between the probability distribution of X and Y. it is also called the Relative entropy two distributions.

$K = \left(P_{X,} P_Y\right) = 0\ iff\ P_X \ = \ P_Y$

and $K = \left(P_{X,} P_Y\right), K = \left(P_{Y,} P_X\right) \ \geq 0$ ___________ (?)

from ***, we can show that (?) holds

$$K = (P_{X,} P_Y = \sum P_X\ (x) \log P_X\ (x) - \ \sum P_X\ (x) \log P_Y\ (x)$$

From Lemma 1 in the note then

$K = \left(P_{X,} P_Y\right) \ \geq 0$

Example 2.2.3

Given $P_X(0) = 0.2$ and $P_X(1) = 0.8$

$P_Y(0) = 0.26$ and $P_Y(1) = 0.74$

then $K(P_{X,}\, P_Y) = P_X(0) \log \frac{P_X(0)}{P_Y(0)} + P_X(1)\ \log \frac{P_X(1)}{P_Y(1)}$

$= \ 0.2\ log_2 \frac{0.2}{0.26} + 0.8\ log_2 \frac{0.8}{0.74}$

$\simeq$ 0.0143 bit

Also,

$K(P_{Y,}\, P_X) = P_Y(0) \log \frac{P_Y(0)}{P_X(0)} + P_Y(1)\ \log \frac{P_Y(1)}{P_X(1)}$

$= \ 0.26\ log_2 \frac{0.26}{0.2} + 0.74\ log_2 \frac{0.74}{0.8}$

$\simeq$ 0.015 bit $\Rightarrow K(P_{X,}\, P_Y) \ \neq \ K(P_{Y,}\, P_X)$

2.3 SOURCE CODING

In a digital communication system, the aim of the designer is to convert any information into digital signal, pass it through the transmission medium, and at the receiving end, reproduce the digital signal exactly. To achieve this objective, two important requirements are;

1. To code any type of information into digital format. Note that the world in analog e.g. voice signals, images etc. There is need to devise mechanism to convert these analog signals into digital format. If the source produces symbols (such as A, B), there is need to convert these symbols into a bit stream. This coding has to be done efficiently so that the smallest number of bits is required for coding.
2. To ensure that the data sent over the channel is not corrupted. Noise introduced on the channels cannot be eliminated, hence there is need to introduce some special coding technique to overcome the effect of noise.

These two aspects have been addressed in source coding theorem and channel coding theorem.

2.3.1 Source Coding Theorem

The source coding theorem states that “the number of bits required to uniquely describe an information source can be approximated to the information content as closely as desired”

Shannon theorem does not state the type of coding technique but puts a limit on the minimum number of bits required. *Engineers are staying to achieve the limit all these 50 > years. Consider a source that produces two symbols A and B with equal probability.

Table 1

Symbol	**Probability**	**Code Word**
A	0.5	1
B	0.5	0

For this coding, we require 1 bit/symbol. Now consider a source that produces these same two symbols, but instead of coding A and B directly, it can be coded as AA, AB, BA, BB. The probabilities of these symbols with their associated codewords are shown here;

Table 2

Symbol	**Probability**	**Code Word**
AA	0.45	0
AB	0.45	10
BA	0.05	110
BB	0.05	111

In this case, there are different sizes of bits or each codeword. So the average number of bits required per symbol can be calculated using the formula

$$L = \sum_{i=1}^{I} P(i)$$

where I = $|X|$

That is, the average or expected length of codeword, denoted by L. where P(i) is the probability and l(i) is the length of the codeword.

from Table 2

L = [1*0.45 + 2*0.45 + 3*0.05 + 3*0.05]

= 1.65 bits/symbol.

The entropy of the source is 1.469 bits/symbol.

So if the source produces the symbols in the following sequence:

A A BA BA AB BB

Then source coding gives the bit stream.

0 110 110 10 111

This coding scheme requires 1.65 bits/symbol. The coding mechanism taking the probabilities into consideration is a better coding technique. The theoretical limit of the number of bits/symbol is the entropy which is 1.469 bits/symbol. The entropy of the source provides the lower bound of the average codeword length of the source code i.e.

$$L\ (c) \geq \frac{H\ (X)}{log_{\,n}}$$

where Pi = n^{-li} ⇒ *L (c) = Hn (x)*

where *n* is the size of the code alphabet.

Proof

$$H\ (x) = -\textstyle\sum_{i=1} Pi \log Pi$$

$$= -\textstyle\sum n^{-li}\ log\ n^{-li}$$

$$= \textstyle\sum lin^{-li}\ (log\ n)$$

$$\log n\ \textstyle\sum li\ n^{-li}$$

$$L\ (C) = \sum_{i=1}^{I} li\ Pi = \sum li\, n^{-li}$$

i.e. $H(x)\ log_n L(c)$

$$\Rightarrow L\ (C) = \frac{H\ (x)}{\log\ n}$$

Only when Pi = n^{-li}

Example 2.3.1: Let X be a r.v. with the following distribution and codeword assignment

$P(X = 1) = ½;$ C (1) = 0

$P(X = 2) = {}^{1}/_{4};$ C (2) = 10

$P(X = 3) = {}^{1}/_{8};$ C (3) = 110

$P(X = 4) = {}^{1}/_{8};$ C (4) = 111

Find (i) the mean information provided by the source X (ii) the expected length of codeword, C (i.e. the average number of bit used.

Solution

(i) The mean information i.e. the entropy, H(x) = $\sum P(xi)log_2 P(xi)$

$$= \frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + 2\ (\frac{1}{8}\log\frac{1}{8})$$

= 1.75 bits/symbol

(ii) the expected length of codeword c, i.e. L (C)

$$L\ (C)\ \sum P(x)\ l(x)$$

$$= \frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{8}(3) + \frac{1}{8}(3)$$

$$= {}^{1}/_{2} + {}^{1}/_{2} + {}^{3}/_{4}$$

= 1.75 bits

$$\Rightarrow \quad L\ (C) \geq \frac{H(x)}{log_2 n}$$

$$\geq \frac{1.75}{log_2 4} \geq \frac{1.75}{2}$$

2.3 ≥ 0.875

2.3.1.1 Classes of Codes

There are basic requirements for a useful symbol code.

First, any encoded string must have a unique coding.

Second, the symbol code must be easy to decode.

Thirdly, the code should achieve as much compression as possible $\left(i.e. L(C)\right)$

(1) Any encoded string must have a unique decoding. A code (x) is said to be non-singular if distinct elements in X map into different strings in F* (where F is the set of finite length strings of symbols from q-ary alphabet, assume F = {0, 1, … q-1} for a q-ary alphabet $(C: X \rightarrow \mathrm{F}*)$).

i.e. for x, y $\begin{smallmatrix} x \rightarrow C\ (x) \\ y \rightarrow C\ (y) \end{smallmatrix}$

that is for any $x\ \neq y,$ where $x, y\ EX$ we have $C(x)\ \neq C(y)$

Non-singularity means that if a single value of X is transmitted, it can be uniquely decoded. The extension of a code C is the mapping from finite length strings of X to finite length strings of F, defined by $C(x_1, x_2, \ldots, x_n = C(x_1), C(x_2), \ldots, (x_n)$

e.g. if C(a) = 00 and C (b) = 11 then

C (aa) = 00 00, C (ab) = 00 11

A code is called uniquely decodable, if its extension is non-singular from the example 2.3.1 with the information given, any sequence of bits can be uniquely decoded e.g 0110111100110 is 134213

(2) The symbol code must be easy to decode

A symbol code is easiest to decode if it is possible to identify the end of a codeword as soon as it arrives which means that no codeword can be a prefix of another codeword. {A word C is a prefix of another word d if a tail string the concatenation ct is identical to d.

e.g. 0 is a prefix of 010 and 011

Any encoded string in a uniquely decodable code has only one possible source string producing it. However, one may have to look at the entire string to determine even the first symbol in the corresponding source string. A code is said to be instantaneously decodable following each code word in any string of codewords can be decoded as soon as its end is reached. A code is called a prefix code or an instantaneous code if no codewords is a prefix of any other codeword.

Example 2.3.1(b)

If C = {0, 10, 110, 111} then the string 010111111010 can be parsed as 0 10 111 110 10

In order to illustrate classes of codes, the table below illustrates the differences.

Table 3: Classes of Codes

	Singular	**Non-singular but not uniquely decodable**	**Uniquely decodable but not instantaneous**	**Instantaneous**
X				
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

Example 2.3.1 (c)

Consider some information source, U, the symbols of which are $U_1, U_2 = 2$ and U_3 and $U_4 = 4$ with the following probability distribution.

U_i	1	2	3	4
$P(U = U_i)$	0.5	0.25	0.125	0.125
$C(U_i)$	0	10	110	111

I. Is this code instantaneously decodable?

II. Encode the message 423312

III. Decode the sequence 101101010

Solution

(i) Any prefix code is instantaneously decodable and conversely any instantaneously decodable code is prefix code. This implies that the code is instantaneously decodable.

(ii) 1423312

011110110110010

(iii) 101101010

= 2322

In order to construct instantaneous codes of minimum expected length to describe a given source, Kraft inequality appears as a sufficient condition.

2.3.1.2 Kraft Inequality Theorem

For any instantaneous code over an alphabet of size, q the codeword length $\ell_1, \ell_2, \ldots \ell_m$ most satisfy the inequality:

$\sum_{i=1}^{m} q^{-\ell i} \leq 1$ ……….. (1) where q = 2 for any binary alphabet

Conversely, given a set o codeword length that satisfy this inequality, J an instantaneous code with these word lengths.

i.e there exists a prefix code C with precisely these codeword lengths, $\ell(x)$.

when equation (1) is satisfied with equality, the corresponding prefix code is called complete code.

For example, using these codewords {0, 10, 11}

$$\sum 2^{-\ell i} = 2^{-1} + 2^{-2} + 2^{-2}$$

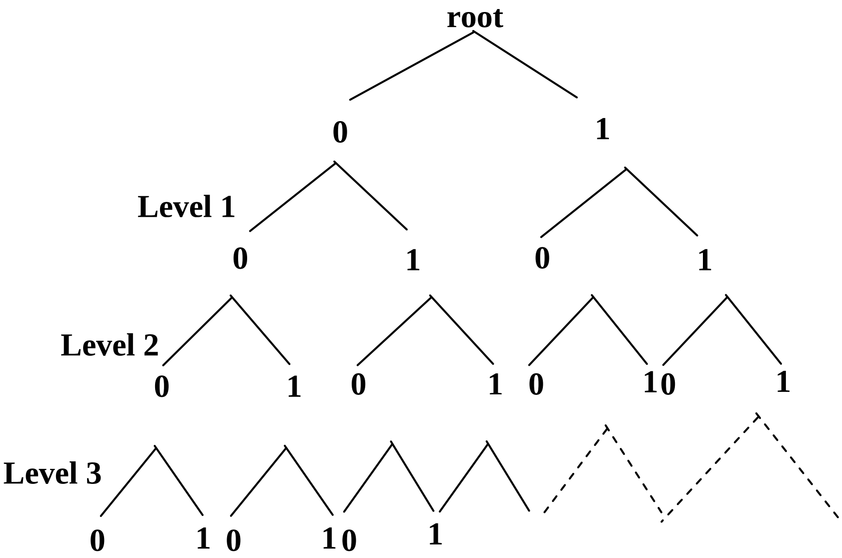
$$= 1/2 + 1/4 + 1/4 = 1$$

$\Rightarrow$ a complete code.

Kraft inequality shows when a prefix code exists and does not answer the question if a given code is indeed prefix-free code.

Proof:

Kraft inequality is proved by building a tree representation of all possible binary strings as

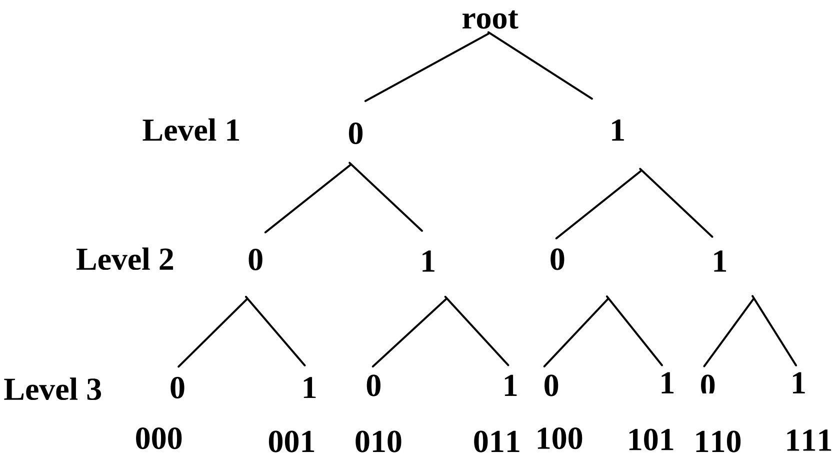


Theorem 1 (Kraft Inequality): Given a message set A. any prefix code C, characterized by a set of codeword len*gth* ℓ *(x); x EA* satisfies the Kraft inequality.

$$\sum_{xEA}\left(\frac{1}{2}\right)\ell^{x} \leq 1$$

Proof:

To prove this inequality, a tree representation of all possible binary strings, is built as



This tree extends downward towards infinity. To each branch point in this tree, a binary string is associated which simply represents all bits one runs into if one follows the directed path from the directed path from the root to that particular branch point e.g. at level 1, we have the branch points 0 and 1, level 2 : 00, 01, 10, 11 and so on.

- At level n, the binary strings associated with the nodes are just the binary representations of the numbers, 0, 1, … $2^n - 1$ e.g. for level 3 it is $2^3 - 1 = 7$ ⇒ binary representation of 0, 1, … 7.
- Those strings that actually occur among the codewords of a prefix code C under consideration are marked with a (in the tree, the codewords occurring up to length 3 are 00, 01 and 1000).

- Since C is a prefix code we know that, once a branch point is selected to correspond to a codeword (marked by a), there can be no other further down along the tree rooted in that particular branch point.
- Weights are associated with the branch points of the tree as follows, staring with weight 1 at the root. The two nodes at level 1 each carry weight ½, the four nodes at level 2 each carry weight ¼ , and so on

  i.e. the 2^K nodes at level K each carry weight 2^{-K}. This construction ensures that irrespective of the level K,

(i) the total weight of all 2^K nodes at that level adds up to 1.

(ii) The total weight of a subset set nodes which together form the complete set of descendants of a node at level $\ell < K$ (containing $2^{K-\ell}$ elements) adds up to $2^{-\ell}$ i.e. the weight of the originating node.

- Now at each level $n \geq 1$, one can divide the set of nodes into three disjoint groups:

(i) nodes that are blocked and may not be used for codewords because they are descendants of a node that has been used as a codeword C*(x)* at a level weight adds up to $2^{-\ell(x)}$

(ii) nodes that are used as codewords at the level n.

(iii) nodes that re unused (and may not represent the first n bits of further codewords of length m > n).

(iv) labeling bitstrings C_n E {0, 1} and recalling that the weight W (C_n) of a node associated with C_n (at level n) is W (C_n) = 2^{-n}, for all $n \geq 1$;

$$\Rightarrow \qquad \sum W\ (Cn) = \underset{C_n;\, blocked}{\Sigma\left(\tfrac{1}{2}\right)n} + \underset{C_n;\, used}{\Sigma\left(\tfrac{1}{2}\right)n} + \underset{C_n;\, unused}{\Sigma\left(\tfrac{1}{2}\right)n}$$

$$\underset{\substack{xEA,\,\ell\,(x)<n \\ \left(\text{of course } \ell\,(\text{x})<n\right)}}{\Sigma\left(\tfrac{1}{2}\right)\ell\,(x)} + \underset{\substack{xEA,\,\ell\,(x)=n \\ \text{(at level 3,we have 3 bitstrings i.e.)}}}{\Sigma\left(\tfrac{1}{2}\right)\ell\,(x)} +$$

By omitting the weight of the unused nodes from the sum, we get

$$\forall \, n \geq 1 : \sum_{xEA,\, \ell\,(x) \leq n} \left(\tfrac{1}{2}\right) \ell\,(x) \leq 1 \quad \text{(of course after removing bloked and unused)}$$

This result follows immediately by taking the limit n → ∞

$$\sum_{xEA} \left(\frac{1}{2}\right)^{(x)} \leq 1$$

Conversely there is another theorem as below:

Theorem 2: If a set of codeword length *ℓ (x)* satisfies the Kraft inequality, there exists a corresponding prefix code having the *ℓ (x)* as codeword lengths.

Proof: This is proved by explicit construction.

We first order the codeword length *ℓ (i)*

i.e. $\ell\,(x_i)$ according to increasing length:

$\ell_1 \leq \ell_2 \leq \ell_3 \leq \ldots\ldots$

Codewords are then constructed in terms of the binary tree drawn earlier (refers)

- Choose for the first codeword C_1 the leftmost branch point at level ℓ_1 in the tree.

  That is, $C_1 =$ 00 … 0 (ℓ_1 zero's)

  Then iterate

- If $\ell_{i+1} = \ell_i$ C_{i+1} is branch point to the right of C_i
- If $\ell_{i+1} > \ell_i$ C_{i+1} is letmost available node at level ℓ_{i+1} (A branchpoint is available if it's not a descendant of a codeword used earlier and these blocked).
- $i \rightarrow i + 1$
- There is a concise algebraic representation of this iterative scheme in terms of arithmetic in base-2 representation viz.

$$C_1 = 0$$

$$C_i + 1 = 2^{(\ell_{i+1} - \ell_i)} \text{ x } (C_i + 1);\ i \geq 1$$

where it is understood that the binary representation of C$_i$ uses ℓ_i bits – allowing as many leading zero's as necessary to achieve this, if C$_i$ > 2^{ℓ_i} – 1

- This construction could only fail, if we allocated a codeword to the right-most node at some level n at a stage where there are still codeword lengths waiting to get a codeword. In terms of the categorization of the codewords at level n, this means that they are all either blocked (by codewords assigned at levels higher up in the tree) or used and no unused ones are left (hence all nodes further down the tree would be blocked), so

$$\sum \left(\frac{1}{2}\right)^{\ell_i} = 1$$

$$i = \ell_i \leq n$$

Example 2.3.1.1(c): To illustrate the construction of refix codes, choose the set

$\{\ell_i\} = \{2, 2, 3, 3, 4, 4\}$

Check for kraft inequality

$$\sum \left(\frac{1}{2}\right)^{\ell_i} = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^4$$

$$= 2\left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4\right]$$

$$= 2\left[\frac{1}{4} + \frac{1}{8} + \frac{1}{16}\right] = \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$$

$$= \frac{7}{8}$$

This satisfies kraft inequality.

This is the construction code

$$C_1 = 0$$
$$C_i + 1 = 2^{(\ell_{i+1} - \ell_i)} \text{ x } (C_i + 1)$$

which generates…

			construction	**result**
			$C_1 \quad = \quad 0$	00
$i = 1$	C_2	=	$2^{(\ell_2 - \ell_1)} \text{ x } C_1 + 1$	
			$= 2^0 * 1 = 1$	01
$i = 2$	C_3	=	$2^{\ell_3 - \ell_2} \text{ x } C_2 + 1$	
			$= 2^{3-2} * 2 = 4$	100
$i = 3$	C_4	=	$2^{\ell_4 - \ell_3} \text{ x } C_3 + 1$	
			$= 2^{3-3} * 4 + 1 = 5$	101
$i = 4$	C_5	=	$2^{\ell_5 - \ell_4} \text{ x } C_4 + 1$	
			$= 2^{4-3} * 5 + 1 = 12$	1100
$i = 5$	C_6	=	$2^{\ell_6 - \ell_5} \text{ x } C_5 + 1$	
			$= 2^{4-4} * 13 = 13$	1101

This is a prefix code, with the required codeword lengths.

Example 2.3.1.1(ii) This example illustrates how the construction fails when the kraft inequality is not satisfied. Choose $\{\ell_i) = \{1, 2, 2, 2\}$.

$$\Sigma({}^{1}\!/_{2})^{\ell i} = {}^{1}\!/_{2} + 3({}^{1}\!/_{2})^{2}$$

$$= {}^{5}\!/_{4} > 1$$

Construction given the correspond code

$C_1 = 0$	0
$C_2 = 2^* C_1 + 1) = 2$	10
$C_3 = C_2 + 1 = 3$	11
$C_4 = C_3 + 1 = 4$	≡ 100 (not a 2 Bit string).

⇒ a prefix code with codeword length $\{\ell_i\}$ violating the kraft inequality cannot be constructed. Once C_3 is found, the rightmost node at level 2 has been reached. (C_3 consists only of 1's); the tentative C_4 constructed via the algebraic. Iteration scheme does not allow a 2-bit representation as required. Another way to see the failure of the construction is to note that C_2 is a prefix of the tentatively constructed C_4!

2.4 Channel Capacity

Shannon introduced the concept of channel capacity, the limit at which data can be transmitted through a medium. The errors in the transmission medium depend on the energy of the signal, the energy of the noise and the bandwidth of the channel. Conceptually, if the bandwidth is high, more data can be pumped in the channel. If the signal energy is high, the effect of noise is reduced.

According to Shannon, the bandwidth of the channel and signal energy and noise energy are related by the formula:

$C = W\ log_2(1 + S/N)$ where

C is channel capacity in bits/second (bps)

W is bandwidth of the channel in H_2

S/N is the signal-to-noise power ratio (SNR)

SNR is measured in dB using the formula:

(S/N)dB = 10 log (Signal power/Noise power)

The value of the channel capacity obtained using this formula is the theoretical maximum.

As an example, consider a voice-grade line for which W = 3100H_2, SNR = 30dB (i.e. the signal-to-noise ratio is 1000:1

$$C = 3100 \log_2 (1 + 1000)$$

$$= 30{,}894 \text{ bps}$$

So, we cannot transmit data at a rate faster than this value in a voice-grade line. An important point to be noted is that in the above formula, Shannon assumed only thermal noise.

To increase C, can W be increased?

No, because increasing W increases noise as well, and SNR will be reduced.

To increase C, can SNR be increased?

No, that results into more noise called intermodulation noise.

SECTION THREE

CHANNEL CODING THEORY

In this lesson, student would learn;

- Error Correction and Detection.
- The basis of coding a discrete information source.
- How a noisy transmission can be formalized by the notion of “channel”.
- The channel capacity and transmission rate
- The fundamental limit for the transmission error in the case of a noisy transmission

3.1 NEED FOR ERROR DETECTION AND CORRECTION

Consider a communication system in which the transmitted bit stream is 1 0 11 0 111 0. The transmitted electrical signal corresponding to this bit stream and the received waveform are shown in Figure 3.1. due to the noise introduced in the transmission medium, the electrical signal is distorted. By using a threshold, the receiver determines whether a 1 is transmitted or a 0 is transmitted. In this case, the receiver decodes the bit stream as 1 0 1 0 0 1 0 1 0.

At two places, the received bit is in error – 1 has become 0 in both places.

2.1 Error Detection

The three widely used techniques for error detection are parity, checksum, and cyclic redundancy check (CRC). Two of these techniques are discussed in the following subsections, refer to your CSC 303 note on discussion on the Parity method.

3.1.1 Block Codes

In block codes, a block of information bits is taken and additional bits are generated. These additional bits are called checksum or cyclic redundancy check (CRC). These two are used for error detection. The procedure used in block coding is shown in Figure 4. The block of information bits (say 8000 bits) and generates additional bits (say 16). The output of the block codes is the original data with the additional 16bits. The additional bits are called checksum or CRC. Block codes can detect errors but cannot correct errors.

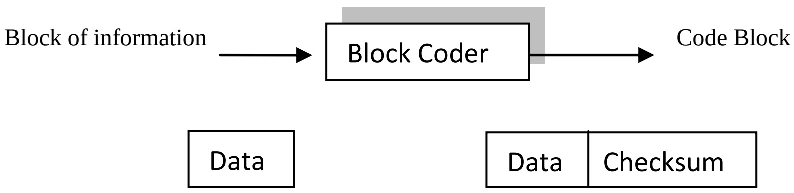


Figure 4: Block Coder

i. CheckSum

Suppose you want to send two characters, C and U. The 7-bit ASCII values for these characters are:

C 1 0 0 0 0 1 1

U 1 0 1 0 1 0 1

In addition to transmitting these bit streams, the binary representation of the sum of these two characters is also sent. The value of C is 67 and the value of U is 85. The sum is 152. The binary representation of 152 is 1 0 0 1 1 0 0 0. This bit stream is also attached to the original binary stream, corresponding to C and U, while transmitting the data. So, the transmitted bit stream is

1 0 0 0 0 1 1 1 0 1 0 1 1 0 0 1 1 0 0 0

At the receiving end, the checksum is again calculated, if the received checksum matches the calculated checksum, then the receiver assumes that the received is OK. The checksum cannot detect all the errors. Also, if the characters are sent in a different order, i.e. if the sequence is changed, the checksum will be the same and hence the receiver assumes that the data is correct.

However, checksum is used mainly because its computation is very easy, and it provides a reasonably good error detection capability.

ii. Cyclic Redundancy Check

CRC is a very powerful technique for detecting errors. Hence, it is extensively used in all data communication systems. Additional bits added to the information bits are called the CRC bits. These bits can be 16 or 32. If the additional bits are 16, the CRC is represented as CRC-16. CRC-32 uses 32 additional bits. Error detection using CRC is very simple. At the transmitting side, CRC is appended to the information bits. At the receiving end, the receiver calculates CRC from the information bits and, if the calculated CRC matches the received CRC, then the receiver knows that the information bits are OK.

Assignment: Download and run a program for calculation of CRC-16 and CRC-32. Modify this program to transmit the message ‘I am a student’.

3.2 Error Correction

If the error rate is high in transmission media such as satellite channels, error-correcting codes are used and have the capability to correct erros. Error correcting codes introduce additional bits, resulting in higher data rate and bandwidth requirements. However, the retransmission can be reduced. Error correcting codes are also called forward-acting error correction (FEC) codes.

Convolutional codes are widely used as error correction codes. The procedure for convolutional coding is shown in Figure 5. The convolutional coder takes a block of information (of n bits) and generate some additional bits (K bits). The additional K bits are derived from the information bits. The output is (n+k) bits. The additional K bits can be used to correct the errors that occurred in the original n bits. n/(n+k) is called rate of the code. For instance, if 2 bits are sent for every 1 information bit, the rate is ½. Then the coding technique is called Rate ½ FEC. If 3 bits are sent for every 2 information bits, the coding technique is called Rate 2/3 FEC. The additional bits are derived from the information bits, and hence redundancy is introduced in the error correcting codes.

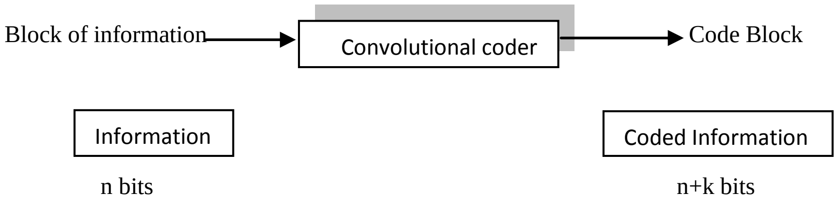


Figure 5: Convolutional Order

In error correcting codes such as convolutional codes, in addition to the information bits, additional redundant bits are transmitted that can be used for error correction at the receiving end. Error correcting codes increase the bandwidth requirement, but they are useful in noisy channel.

For example, in many radio systems, error rate is very high, and so FEC is used. In Bluetooth radio systems, Rate 1/3 FEC is used. In this scheme, each bit is transmitted three times. To transmit bits b0b1b2b3, the actual bits transmitted using Rate 1/3 FEC are b0b0b0b1b1b1b2b2b2b3b3b3

At the receiver, error correction is possible. If the received bit stream is 101000111000111, it is very easy for the receiver to know that the second bit is received in error, and it can be corrected.

A number of FEC coding schemes have been proposed that increase the delay in processing and also the bandwidth requirement but help in error correction. Shannon laid the foundation for channel coding, and during the last five decades, hundreds of error-correcting codes have been developed.

3.3 Coding Discrete Information Source

When dealing with "information", one of the basic goals is to transmit it reliably. In this context, "transmit" both means "transmitting some information from one point to another", as we usually understand it, but also to "transmit" it through time; for instance to store it somewhere (to memorize it) and then retrieve it later on. In both cases however, transmission of information can, in real life, hardly be achieved in a fully reliable manner. There always exists a risk of distortion of the transmitted information: some noise on the line, some leak of the memory or the hard disk storing the information, etc.

What effect does noise have on the transmission of messages? Several situations could be possible:

- it is never possible to transmit any messages reliably (too much noise);
- it is possible to transmit messages with a "reasonable" error probability;
- it is possible to transmit messages with an error probability which is as small as we can wish for (using error correcting codes).

The purpose of the present section is to study how coding can help transmitting information in a reliable way, even in the presence of noise during the transmission. The basic idea of such codings is to try to add enough redundancy in the coded message so that transmitting it in "reasonably" noisy conditions leaves enough information undisturbed for the receiver to be able to reconstruct the original message without distortion.

Of course, the notions of "enough redundancy" and "reasonably noisy conditions" need to be specified further; and even quantified and related. This will be done by first formalizing a bit further the notion of "noisy transmission", by introducing the notion of a "communication channel", which is addressed in the next sub section

As we will see in section 4.3, the two fundamental notions ruling noisy transmissions
are the channel capacity and the rate use for transmitting the symbols of the messages.

3.3.1 Communication Channels

Roughly speaking, a communication channel (shorter "channel") represents all that could happen to the transmitted messages between their emission and their reception.

A message is a sequence of symbols. A symbol is simply an element of a set, called an alphabet. In this course, only finite alphabets will be addressed. The input sequence X_1,X_2,X_3, . . . (i.e. the message to be transmitted) is fully determined by the source alone; but the transmission determines the resulting conditional probabilities of the output sequence Y_1, Y_2, Y_3, . . . (i.e. the message received) knowing the input sequence.

In mathematical terms, the channel specifies the conditional probabilities of the various messages that can be received, conditionally to the messages that have been emitted;

i.e. $P(Y_1 = y_1, ..., Y_n = y_n|X_1 = x_1, ...,X_n = x_n)$ for all possible n and values

y_1, x_1, ..., y_n, x_n.

Definition 1 (Discrete Memoryless Channel)

The discrete memory- less channel (DMC) is the simplest kind of communication channel. Formally, DMC consists of three quantities:

1. a discrete input alphabet, V_X, the elements of which represent the possible emitted symbols for all input messages (the source X);

2. a discrete output alphabet, V_Y , the elements of which represent the possible received symbols (output sequence); and

3. for each $x \in V_X$, the conditional probability distributions $p_{Y\,|X=x}$ over V_Y which describe the channel behavior in the manner that, for all n = 1, 2, 3, . . .:

$$P(Y_n = y_n|X_1 = x_1, . . . , X_n = x_n, Y_1 = y_1, . . . , Y_{n-1} = y_{n-1})$$
$$= P(Y = y_n|X = x_n) \qquad \text{eqn 3.1}$$

These are called the transmission probabilities of the channel.

Equation (3.1) is the mathematical statement that corresponds to the “memoryless” nature of the DMC. What happens to the signal sent on the n-th use of the channel is independent of what happens on the previous n − 1 uses.

Notice also that (3.1) implies that the DMC is time-invariant, since the probability distribution $p_{Yn|xn}$ does not depend on n.

When V_X and V_Y are finite, a DMC is very often specified by a diagram where:

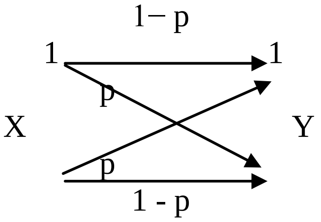
1. the nodes on the left indicate the input alphabet V_X;
2. the nodes on the right indicate the output alphabet V_Y; and
3. the directed branch from x_i to y_j is labeled with the conditional probability $p_{Y\,|X=xi}(y_j)$ (unless this probability is 0, in which case the branch is simply omitted.)

Example 3.3 (a) (Binary Symmetric Channel)

The simplest (non trivial) case of DMC is the binary symmetric channel (BSC), for which $V_X = V_Y = \{0, 1\}$ (“binary”) and $p_{Y\,|X=0}(1) = p_{Y\,|X=1}(0)$ (“symmetric”).

This value $p = p_{Y\,|\,X\,=\,0}(1) = p_{Y\,|X=1}(0)$ is called the error rate and is the only parameter of the BSC. Indeed, $p_{Y\,|X=0}(0) = p_{Y\,|X=1}(1) = 1 - p$.

The BSC is represented by the following diagram:



Example 3.3 (b) (Noisy Transmission over a Binary Symmetric Channel)

Suppose we want to transmit the 8 following messages: 000, 001, 010, 011, 100, 101, 110 and 111.

Suppose that the channel used for transmission is noisy in such a way that it changes one symbol over ten, regardless of everything else; i.e. each symbol has a probability $p = 0.1$ to be ”flipped” (0 into 1, and 1 into 0). Such a channel is a BSC with an error rate equal to $p = 0.1$, What is the probability to transmit one of our messages correctly?

Regardless of which message is sent, this probability is

$$(1 - p)^3 = 0.93 = 0.719$$

(corresponding to the probability to transmit 3 times one bit without error).

Therefore, the probability to receive and erroneous message is 0.281, i.e 28%; which is quite high!

Suppose now we decide to code each symbol of the message by twice itself:

message	000	001	010	011	100	…	111
Code	000000	000011	001100	001111	110000	…	111111

What is now the probability to have a message sent correctly? In the same way, this is

$$(1 - p)^6 = 0.531$$

And the probability to receive and erroneous message is now 0.469... ...worse than previously, it seems!

However, what is the probability to receive an erroneous message which seems to be valid; i.e. what is the probability to receive a erroneous message and to not detect it as wrong? Not detecting an erroneous message means that two corresponding symbol have both been changed. If for instance we sent 000000, but 110000 is received, there is not way to see that some errors occurred. However, if 010000 is received, clearly at least one error occurred (and retransmission could for instance be required).

So, not detecting an error could come either from 2 changes (at the corresponding places) or 4 changes or the whole 6 symbols. What is the probability to change 2 symbols?

Answer:

$$\binom{6}{2} p^2(1-p)^4 = 15p^2(1-p)^4$$

What is the probability to change 2 corresponding symbols? Only

$$\binom{3}{1} p^2(1-p)^4 = 3p^2(1-p)^4$$

Similarly, the probability to change 4 corresponding symbols is $3p^4(1-p)^2$, and to change the whole six symbols is p^6.

Therefore, the probability of not detecting an error is

$3p^2(1-p)^2$ + $3p^4(1-p)^2$ + p^6 = 0.020 which is much smaller.

This means that the probability to make a error in the reception (i.e. to trust an erroneous message without being aware of) is only 0.02.

Conclusion: some codings are better than other for the transmission of messages over a noisy channel.

3.4 Channel Capacity

The purpose of a channel is to transmit messages ("information") from one point (the input) to another (the output). The channel capacity_ precisely measures this ability: it is the maximum average amount of information the output of the channel can bring on the input.

Recall that a DMC if fully specified by the conditional probability distributions $p_{Y|X=x}$ (where *X* stands for the input of the channel and *Y* for the output). The input probability distribution $p_X(x)$ is not part of the channel, but only of the input source used. The capacity of a channel is thus defined as the maximum mutual information I(X; Y) that can be obtained among all possible choice of $p_{X(x).}$ More formally, The capacity C of a Discrete Memoryless Channel is defined as

$$C = \max_{P_x} I(X; Y)$$

where *X* stands for the input of the channel and *Y* for the output.

Example 3.4 (Capacity of BSC) What is the capacity C of a BSC, defined in example 3.3?

First notice that, by definition of mutual information,

$$C = \max\,(H(Y) - H(Y\,|X))$$

Furthermore, since in the case of a BSC, $P(Y \neq X) = p$ and $P(Y = X) = 1 - p$, we have $H(Y\,|X) = -p \log(p) - (1-p) \log(1-p) =: h'(p)$, which does not depend on p_X. Therefore,

$$C = \max_{P_x}\,(H(Y)) - h'(p)$$

Since Y is a binary random variable, we have: $H(Y\,) \leq \log 2$, i.e. $H(Y) \leq 1$ bit.

Can this maximum be reached for some p_X? Indeed, yes: if X is uniformly distributed, we have $p_Y\,(0) = p{\cdot}p_X(1) + (1-p){\cdot}p_X(0) = 0.5\,p + 0.5\,(1-p) = 0.5$; which means that Y is also uniformly distributed, leading to $H(Y\,) = 1$ bit.

Therefore: $\max_X H(Y\,) = 1$ bit and

$$C = 1 - h'(p) \quad \text{(in bits)}$$

SECTION FOUR

MODULATION

In this lesson, student would learn;

- The need for modulation.
- Analog Modulation and digital Modulation.
- Analog Modulation Techniques.
- Digital Modulation Techniques

4.1 What is Modulation?

Modulation can be defined as superimposition of the signal containing the information on a high-frequency carrier. If we have to transmit voice that contains frequency components up to 4kHz, we superimpose the voice signal on a carrier of, say, 140MHz. the input voice signal is called the modulating signal. The transformation of superimposition is called the modulation. The hardware that carries out this transformation is called the MODULATOR. The output of the modulator is called the modulated signal.

Suppose we want to transmit two voice channels from one place to another. If we combine the two voice signals and transmit them on the medium, it is impossible to separate the voice conversations at the receiving end. This because both voice channels occupy the same frequency band, 300Hz to about 4kHz. A better way of transmitting the two voice channels is to put them in different frequency bands and then send them, translating the voice channels into different bands.

Low frequency signals have poor radiation capability and so low-frequency signals such as voice signals are translated into high frequencies. There is need to superimpose the voice signal onto a high-frequency signal to transmit to transmit over large distances. This high-frequency signal is called the carrier, and the modulation is called *carrier modulation.*

When different voice signals are modulated to different frequencies, we can transmit all these modulated signals together. There will be no interference.

If radio is used as the transmission medium, the radio signal has to be sent through an antenna. The size of the antenna decreases as the frequency of the signal goes up. If the voice is transmitted without superimposing it on a high-frequency carrier, the antenna size should be 5,000 meters.

For these reasons, modulation is an important transformation of the signal that is used in every communication system.

To summarize, modulation allows:

- Transmitting signals over large distances, because low-frequency signals have poor radiation characteristics.
- It is possible to combine a number of baseband signals and send them through the medium, provided different carrier frequencies are used for different baseband signals.
- Small antennas can be used if radio is the transmission medium.

4.1.1 Types of Modulation

In general modulation can be defined as transformation of a signal. Carrier modulation transforms a carrier in such a way that the transformed carrier contains the information of the modulating signal.

Many carrier modulation techniques have been proposed in the literature. The most fundamental techniques that are used extensively in both analog and digital communication system would be studied in this class. Carrier modulation can be broadly divided into two categories:

- Analog Modulation
- Digital Modulation

4.2 Analog Modulation

Analog modulation is used extensively in broadcasting of audio and video programs, and many old telecommunication systems are also based on analog modulation. All newly developed systems use digital modulation. For broadcasting, analog modulation continues to play a vital role.

The various analog modulation techniques are:

- Amplitude Modulation (AM)
- Frequency Modulation (FM)
- Phase Modulation (PM)

FM and PM are known as angle modulation techniques.

4.3 Digital Modulation

The three important digital modulation techniques are

- Amplitude Shift Keying (ASK)
- Frequency Shift Keying (FSK)
- Phase Shift Keying (PSK)

READING LIST

Prasad, K. V. (2004) “Principles of Digitaal Communication Systems and Computer Networks“, Charles River Media, INC, Hingham, Massachusetts.

F. Bavaud J.-C. Chappelier J. Kohlas ” An Introduction to Information Theory and Applications” Version 2.04