INFORMATION THEORY

ER. FARUK BIN POYEN, Asst. Professor

DEPT. OF AEIE, UIT, BU, BURDWAN, WB, INDIA

faruk.poyen@gmail.com

Contents..

- Information Theory
- What is Information?
- Axioms of Information
- Information Source
- Information Content of a Discrete Memoryless Source
- Information Content of a Symbol (i.e. Logarithmic Measure of Information)
- Entropy (i.e. Average Information)
- Information Rate
- The Discrete Memoryless Channels (DMC)
- Types of Channels
- Conditional And Joint Entropies
- The Mutual Information
- The Channel Capacity

Contents: Contd..

- Capacity of an Additive Gaussian Noise (AWGN) Chanel - Shannon – Hartley Law
- Channel Capacity: Detailed Study – Proof of Shannon – Hartley Law expression.
- The Source Coding
- Few Terms Related to Source Coding Process
- The Source Coding Theorem
- Classification of Codes
- The Kraft Inequality
- The Entropy Coding – Shannon – Fano Coding; Huffman Coding.
- Redundancy
- Entropy Relations for a continuous Channel
- Analytical Proof of Shannon – Hartley Law
- Rate Distortion Theory
- Probability Theory

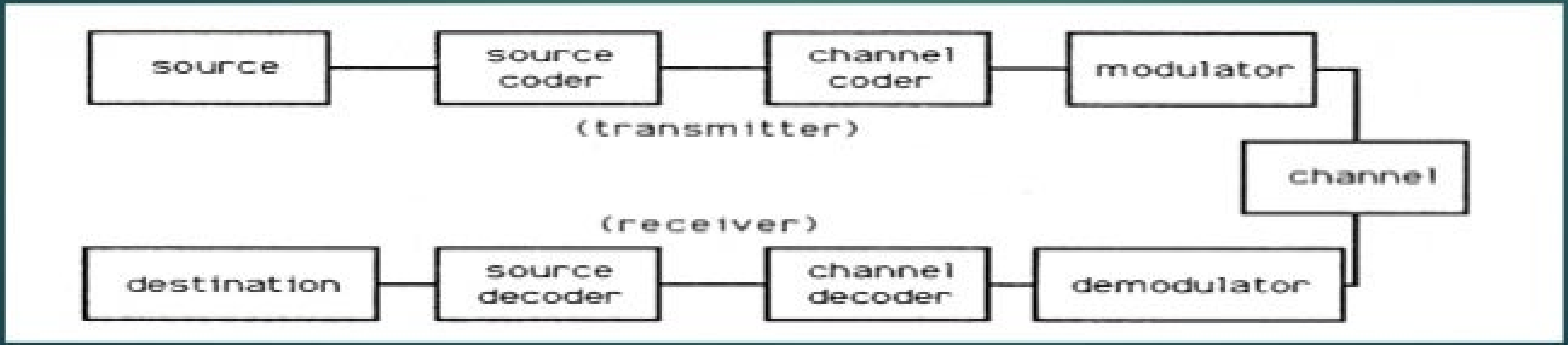
Introduction:

- The purpose of communication system is to carry information bearing base band signals from one place to another placed over a communication channel.
- Information theory is concerned with the fundamental limits of communication.
- What is the ultimate limit to data compression?
- What is the ultimate limit of reliable communication over a noisy channel, e.g. how many bits can be sent in one second over a telephone line?
- Information Theory is a branch of probability theory which may be applied to the study of the communication systems that deals with the mathematical modelling and analysis of a communication system rather than with the physical sources and physical channels.
- Two important elements presented in this theory are Binary Source (BS) and the Binary Symmetric Channel (BSC).
- A binary source is a device that generates one of the two possible symbols ‘0’ and ‘1’ at a given rate ‘r’, measured in symbols per second.

- These symbols are called bits (binary digits) and are generated randomly.
- The BSC is a medium through which it is possible to transmit one symbol per time unit. However this channel is not reliable and is characterized by error probability ‘p’ $(0 \leq p \leq 1/2)$ that an output bit can be different from the corresponding input.
- Information theory tries to analyse communication between a transmitter and a receiver through an unreliable channel and in this approach performs an analysis of information sources, especially the amount of information produced by a given source, states the conditions for performing reliable transmission through an unreliable channel.
- The source information measure, the channel capacity measure and the coding are all related by one of the Shannon theorems, the channel coding theorem which is stated as: ‘If the information rate of a given source does not exceed the capacity of a given channel then there exists a coding technique that makes possible transmission through this unreliable channel with an arbitrarily low error rate.’

- There are three main concepts in this theory:

1. The first is the definition of a quantity that can be a valid measurement of information which should be consistent with a physical understanding of its properties.
2. The second concept deals with the relationship between the information and the source that generates it. This concept will be referred to as the source information. Compression and encryptions are related to this concept.
3. The third concept deals with the relationship between the information and the unreliable channel through which it is going to be transmitted. This concept leads to the definition of a very important parameter called the channel capacity. Error - correction coding is closely related to this concept.



What is Information?

- Information of an event depends only on its probability of occurrence and is not dependent on its content.
- The randomness of happening of an event and the probability of its prediction as a news is known as information.
- The message associated with the least likelihood event contains the maximum information.

Axioms of Information:

1. Information is a non-negative quantity: $I(p) \geq 0$.
2. If an event has probability 1, we get no information from the occurrence of the event: $I(1) = 0$.
3. If two independent events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two information: $I(p_1 * p_2) = I(p_1) + I(p_2)$.
4. $I(p)$ is monotonic and continuous in p.

Information Source:

- An information source may be viewed as an object which produces an event, the outcome of which is selected at random according to a probability distribution.
- The set of source symbols is called the **source alphabet** and the elements of the set are called **symbols** or **letters**.
- Information source can be classified as having memory or being memoryless.
- A source with memory is one for which a current symbol depends on the previous symbols.
- A memoryless source is one for which each symbol produced is independent of the previous symbols.
- *A discrete memoryless source (DMS) can be characterized by the list of the symbol, the probability assignment of these symbols and the specification of the rate of generating these symbols by the source.*

Information Content of a DMS:

- The amount of information contained in an event is closely related to its uncertainty.
- A mathematical measure of information should be a function of the probability of the outcome and should satisfy the following axioms:

a) Information should be proportional to the uncertainty of an outcome

b) Information contained in independent outcomes should add up.

Information Content of a Symbol (i.e. Logarithmic Measure of Information):

- Let us consider a DMS denoted by ‘x’ and having alphabet {x1, x2, ……, xm}.
- The information content of the symbol xi, denoted by I (x_i) is defined by
- $I(x_i) = \log_b \frac{1}{P(x_i)} = -\log_b P(x_i)$
- where P(x_i) is the probability of occurrence of symbol x_i.
- For any two independent source messages xi and xj with probabilities P_i and P_j respectively and with joint probability P (x_i, x_j) = Pi Pj, the information of the messages is the addition of the information in each message. $I_{ij} = I_i + I_j$.

- Note that $I(x_i)$ satisfies the following properties.

1. $I(x_i) = 0$ for $P(x_i) = 1$
2. $I(x_i) \geq 0$
3. $I(x_i) > I(x_j)$ if $P(x_i) < P(x_j)$
4. $I(x_i, x_j) = I(x_i) + I(x_j)$ if x_i and x_j are independent

- **Unit of I (x_i):** The unit of I (x_i) is the bit (binary unit) if b = 2, Hartley or decit if b = 10 and nat (natural unit) if b = e. it is standard to use b = 2.

$$\log_2 a = \ln a/\ln 2 = \log a/\log 2$$

Entropy (i.e. Average Information):

- Entropy is a measure of the uncertainty in a random variable. The entropy, H, of a discrete random variable X is a measure of the amount of uncertainty associated with the value of X.
- For quantitative representation of average information per symbol we make the following assumptions:

i) The source is stationary so that the probabilities may remain constant with time.

ii) The successive symbols are statistically independent and come from the source at an average rate of ‘r’ symbols per second.

- The quantity H(X) is called the entropy of source X. it is a measure of the average information content per source symbol.
- The source entropy H(X) can be considered as the average amount of uncertainty within the source X that is resolved by the use of the alphabet.
- $H(X) = E\ [I(x_i)] = -\ \Sigma P(x_i)\ I(x_i) = -\ \Sigma P(x_i) \log_2 P(x_i)$ b/symbol.

- **Entropy for Binary Source:**

$$H(X) = -{}^{1}/_{2}\log_2 {}^{1}/_{2} - {}^{1}/_{2}\log_2 {}^{1}/_{2} = 1\ b/symbol$$

- The source entropy H(X) satisfies the relation: $\mathbf{0 \leq H(X) \leq \log_2 m}$, where m is the size of the alphabet source X.

- **Properties of Entropy:**

1. $0 \leq H\,(X) \leq \log_2 m$; m = no. of symbols of the alphabet of source X.
2. When all the events are equally likely, the average uncertainty must have the largest value i.e. $\log_2 m \geq H\,(X)$
3. H (X) = 0, if all the $P(x_i)$ are zero except for one symbol with P = 1.

Information Rate:

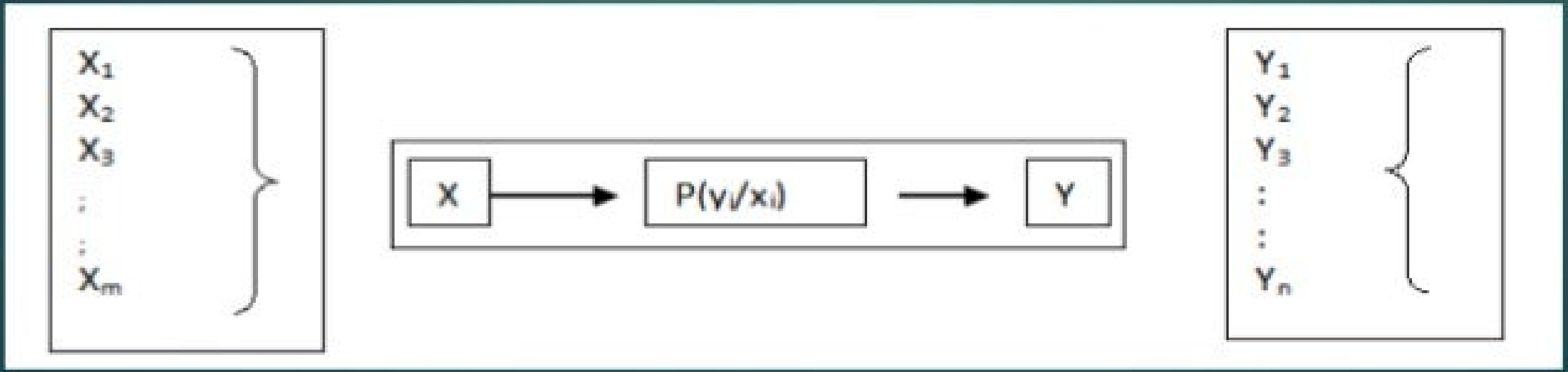
- If the time rate at which X emits symbols is ‘r’ (symbols s), the information rate R of the source is given by
- **R = r H(X) b/s** [(symbols / second) * (information bits/ symbol)].
- R is the information rate. H(X) = Entropy or average information.

The Discrete Memoryless Channels (DMC):

1. **Channel Representation:** A communication channel may be defined as the path or medium through which the symbols flow to the receiver end.

A DMC is a statistical model with an input X and output Y. Each possible input to output path is indicated along with a conditional probability P (y_j/x_i), where P (y_j/x_i) is the conditional probability of obtaining output y_j given that the input is x_1 and is called a **channel transition probability.**

1. A channel is completely specified by the complete set of transition probabilities. The channel is specified by the matrix of transition probabilities [P(Y/X)]. This matrix is known as **Channel Matrix**.



$$[P(Y/X)] = \begin{bmatrix} P(y_1/x_1) & \cdots & P(y_n/x_1) \\ \vdots & \ddots & \vdots \\ P(y_1/x_m) & \cdots & P(y_n/x_m) \end{bmatrix}$$

Since each input to the channel results in some output, each row of the column matrix must sum to unity. This means that

$$\sum_{j=1}^{n} P(y_j/x_i) = 1 \; for \; all \; i$$

Now, if the input probabilities P(X) are represented by the row matrix, we have

$$[P(X)] = [P(x_1)P(x_2) \ldots . P(x_m)]$$

- Also the output probabilities P(Y) are represented by the row matrix, we have

$$[P(Y)] = [P\left(y_1\right)P\left(y_2\right) \ldots . P(y_n)]$$

Then

$$[P(Y)] = [P(X)][P(Y/X)]$$

Now if P(X) is represented as a diagonal matrix, we have

$$[P(X)]_d = \begin{bmatrix} P(x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P(x_m) \end{bmatrix}$$

Then

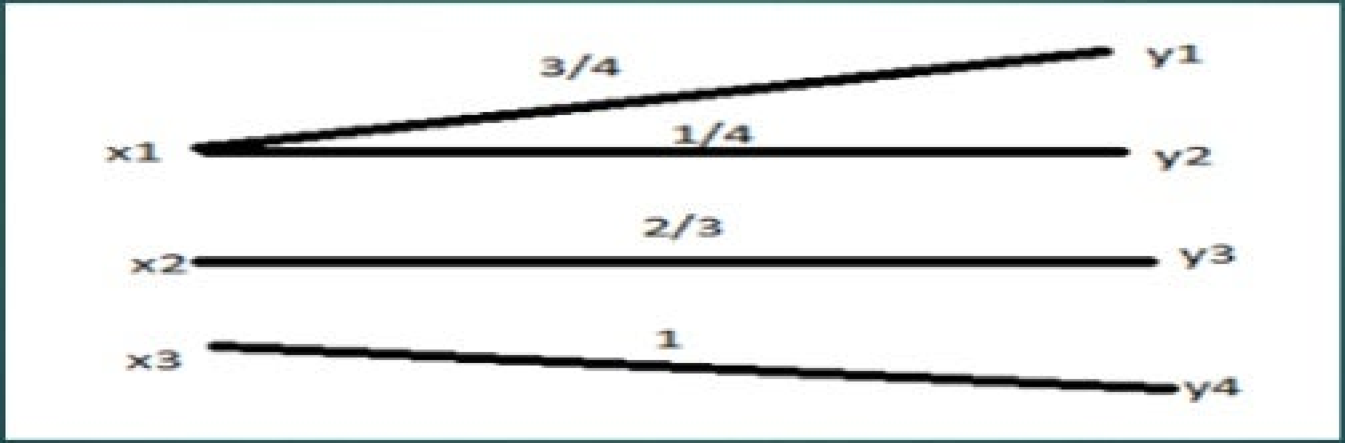
$$[P(X,Y)] = [P(X)]_d[P(Y/X)]$$

- Where the (i, j) element of matrix [P(X,Y)] has the form P(x_i, y_j).
- The matrix [P(X, Y)] is known as the ***joint probability matrix*** and the element P(x_i, y_j) is the joint probability of transmitting x_i and receiving y_j.

Types of Channels:

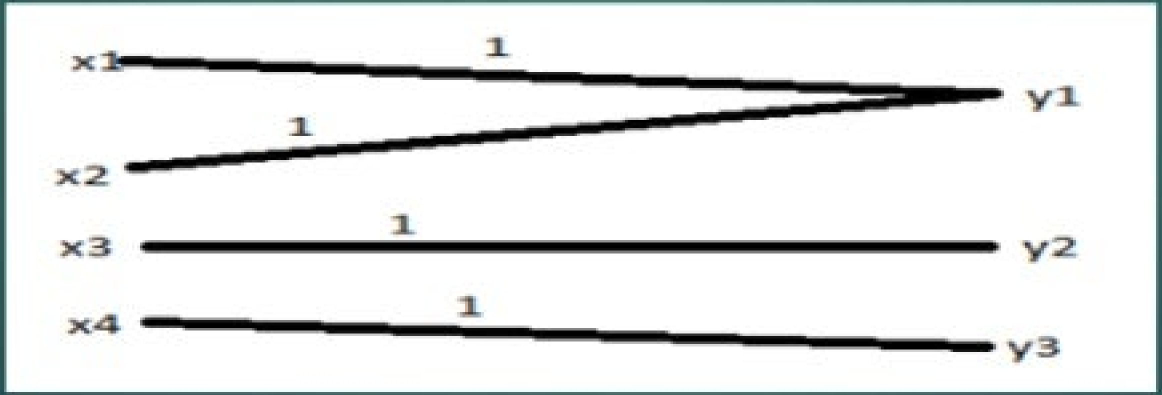
- Other than discrete and continuous channels, there are some special types of channels with their own channel matrices. They are as follows:
- **Lossless Channel:** A channel described by a channel matrix with only one non – zero element in each column is called a lossless channel.

$$\left[P\left(\frac{Y}{X}\right)\right] = \begin{bmatrix} {}^{3}\!/_{4} & {}^{1}\!/_{4} & 0 & 0 \\ 0 & 0 & {}^{2}\!/_{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



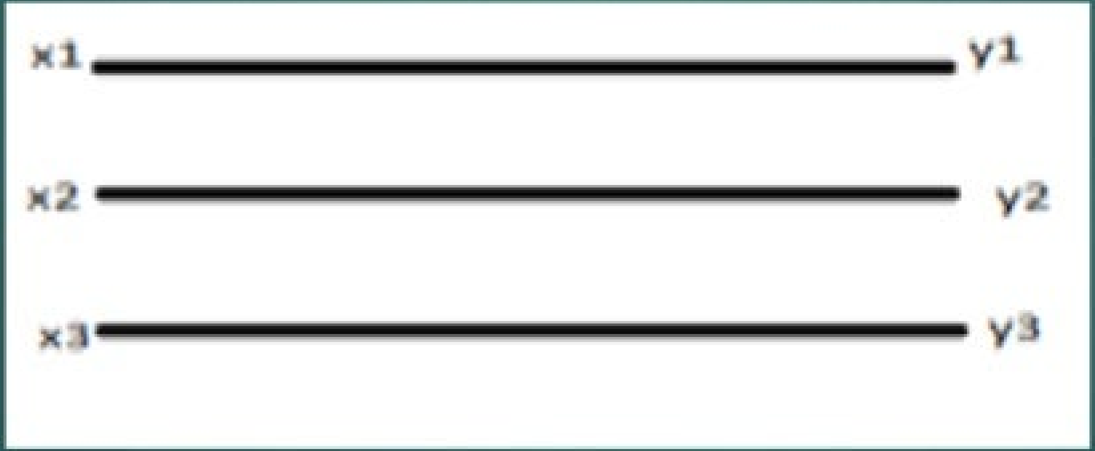
- **Deterministic Channel:** A channel described by a channel matrix with only one non – zero element in each row is called a deterministic channel.

$$\left[P\left(\frac{Y}{X}\right)\right] = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



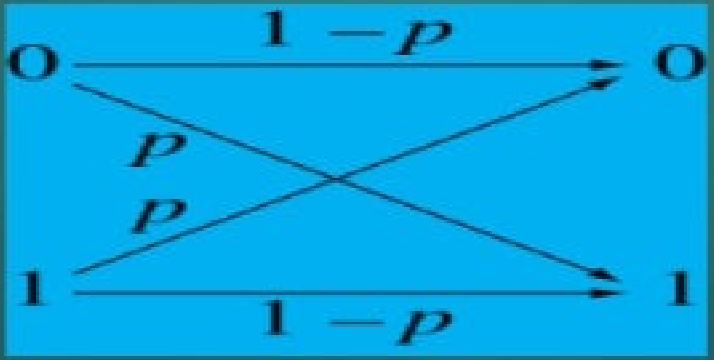
- **Noiseless Channel:** A channel is called noiseless if it is both lossless and deterministic. For a lossless channel, m = n.

$$\left[P\left(\frac{Y}{X}\right)\right] = \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$



- **Binary Symmetric Channel:** BSC has two inputs ($x_1 = 0$ and $x_2 = 1$) and two outputs ($y_1 = 0$ and $y_2 = 1$).
- This channel is symmetric because the probability of receiving a 1 if a 0 is sent is the same as the probability of receiving a 0 if a 1 is sent.

$$\left[P\left(\frac{Y}{X}\right)\right] = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$



Conditional And Joint Entropies:

- Using the input probabilities P (x_i), output probabilities P (y_j), transition probabilities P (y_j/x_i) and joint probabilities P (x_i, y_j), various entropy functions for a channel with m inputs and n outputs are defined.

$$H(X) = -\sum_{i=1}^{m} \mathrm{P}(\mathrm{x_i}) \log_2 \mathrm{P}(\mathrm{x_i})$$

$$H(Y) = -\sum_{j=1}^{n} \mathrm{P}(\mathrm{y_j}) \log_2 \mathrm{P}(\mathrm{y_j})$$

$$H\left(\frac{X}{Y}\right) = -\sum_{j=1}^{n}\sum_{i=1}^{m} P(x_i, y_j) \log_2 (x_i/y_j)$$

$$H\left(\frac{Y}{X}\right) = -\sum_{j=1}^{n}\sum_{i=1}^{m} P(x_i, y_j) \log_2 (y_j/x_i)$$

$$H(X,Y) = -\sum_{j=1}^{n}\sum_{i=1}^{m} P(x_i, y_j) \log_2 (x_i, y_j)$$

- $H(X)$ is the average uncertainty of the channel input and $H(Y)$ is the average uncertainty of the channel output.
- The conditional entropy $H(X/Y)$ is a measure of the average uncertainty remaining about the channel input after the channel output has been observed. $H(X/Y)$ is also called equivocation of X w.r.t. Y.
- The conditional entropy $H(Y/X)$ is the average uncertainty of the channel output given that X was transmitted.
- The joint entropy $H(X, Y)$ is the average uncertainty of the communication channel as a whole. Few useful relationships among the above various entropies are as under:
- $H(X, Y) = H(X/Y) + H(Y)$
- $H(X, Y) = H(Y/X) + H(X)$
- $H(X, Y) = H(X) + H(Y)$
- $H(X/Y) = H(X, Y) – H(Y)$
- X and Y are statistically independent.

- The **conditional entropy** or conditional uncertainty of X given random variable Y (also called the equivocation of X about Y) is the average conditional entropy over Y.
- The **joint entropy** of two discrete random variables X and Y is merely the entropy of their pairing: (X, Y), this implies that if X and Y are independent, then their joint entropy is the sum of their individual entropies.

The Mutual Information:

- Mutual information measures the amount of information that can be obtained about one random variable by observing another.
- It is important in communication where it can be used to maximize the amount of information shared between sent and received signals.
- The mutual information denoted by I (X, Y) of a channel is defined by:

$$I(X; Y) = H(X) - H\left(\frac{X}{Y}\right) b/symbol$$

- Since H (X) represents the uncertainty about the channel input before the channel output is observed and H (X/Y) represents the uncertainty about the channel input after the channel output is observed, the mutual information I (X; Y) represents the uncertainty about the channel input that is resolved by observing the channel output.

- **Properties of Mutual Information I (X; Y)**
- **I (X; Y) = I(Y; X)**
- **I (X; Y) ≥ 0**
- **I (X; Y) = H (Y) – H (Y/X)**
- **I (X; Y) = H(X) + H(Y) – H(X,Y)**

- The Entropy corresponding to mutual information [i.e. I (X, Y)] indicates a measure of the information transmitted through a channel. Hence, it is called **‘Transferred information’.**

The Channel Capacity:

- The channel capacity represents the maximum amount of information that can be transmitted by a channel per second.
- To achieve this rate of transmission, the information has to be processed properly or coded in the most efficient manner.
- **Channel Capacity per Symbol C_S:** The channel capacity per symbol of a discrete memoryless channel (DMC) is defined as

$$C_S = \max_{\{P(x_i)\}} I(X;Y) b/symbol$$

Where the maximization is over all possible input probability distributions $\{P(x_i)\}$ on X.

- **Channel Capacity per Second C:** I f 'r' symbols are being transmitted per second, then the maximum rate od transmission of information per second is 'r C_S'. this is the channel capacity per second and is denoted by C (b/s) i.e.

$$C = rC_S \, b/s$$

Capacities of Special Channels:

- **Lossless Channel:** For a lossless channel, H (X/Y) = 0 and I (X; Y) = H (X).
- Thus the mutual information is equal to the input entropy and no source information is lost in transmission.

$$C_S = \max_{\{\mathrm{P(x_i)}\}} H(X) = \log_2 m$$

Where m is the number of symbols in X.

- **Deterministic Channel:** For a deterministic channel, H (Y/X) = 0 for all input distributions P (x_i) and I (X; Y) = H (Y).
- Thus the information transfer is equal to the output entropy. The channel capacity per symbol will be

$$C_S = \max_{\{\mathrm{P(x_i)}\}} H(Y) = \log_2 n$$

where n is the number of symbols in Y.

- **Noiseless Channel:** since a noiseless channel is both lossless and deterministic, we have I (X; Y) = H (X) = H (Y) and the channel capacity per symbol is

$$C_S = \log_2 m = \log_2 n$$

- **Binary Symmetric Channel:** For the BSC, the mutual information is

$$I\,(X;Y) = H\,(Y) + p\log_2 p + (1-p)\log_2 (1-p)$$

And the channel capacity per symbol will be

$$C_S = 1 + p\log_2 p + (1-p)\log_2 (1-p)$$

Capacity of an Additive Gaussian Noise (AWGN) Channel - Shannon – Hartley Law

- The Shannon – Hartley law underscores the fundamental role of bandwidth and signal – to – noise ration in communication channel. It also shows that we can exchange increased bandwidth for decreased signal power for a system with given capacity C.

- In an additive white Gaussian noise (AWGN) channel, the channel output Y is given by

$$Y = X + n$$

- Where X is the channel input and n is an additive bandlimited white Gaussian noise for zero mean and variance σ^2.

- The capacity C_S of an AWGN channel is given by

$$C_s = \max_{\{P(x_i)\}} I(X;Y) = {}^{1}/_{2} \log_2\left(1 + {}^{S}/_{N}\right) \text{ b/sample}$$

- Where S/N is the signal – to – noise ratio at the channel output.
- If the channel bandwidth B Hz is fixed, then the output y(t) is also a bandlimited signal completely characterized by its periodic sample values taken at the Nyquist rate 2B samples/s.

- Then the capacity C (b/s) of the AWGN channel is limited by

$$C = 2B * C_S = B \log_2\left(1 + {}^{S}/_{N}\right)\ b/s$$

- This above equation is known as the **Shannon – Hartley Law.**

Channel Capacity: Shannon – Hartley Law Proof.

- The bandwidth and the noise power place a restriction upon the rate of information that can be transmitted by a channel. Channel capacity C for an AWGN channel is expressed as

$$C = B\log_2(1 + {}^{S}/_{N})$$

- Where B = channel bandwidth in Hz; S = signal power; N = noise power;
- **Proof:** Assuming signal mixed with noise, the signal amplitude can be recognized only within the root mean square noise voltage.
- Assuming average signal power and noise power to be S watts and N watts respectively the RMS value of the received signal is $\sqrt{(S+N)}$ and that of noise is $\sqrt{N}$.
- Therefore the number of distinct levels that can be distinguished without error is expressed as

$$M = {}^{\sqrt{S+N}}\Big/_{\sqrt{N}} = \sqrt{1 + {}^{S}/_{N}}$$

- The maximum amount of information carried by each pulse having $\sqrt{1 + {}^{S}\!/_{N}}$ distinct levels is given by

$$I = \log_2 \sqrt{1 + {}^{S}\!/_{N}} = {}^{1}\!/_{2} \log_2(1 + {}^{S}\!/_{N})\ bits$$

- The channel capacity is the maximum amount of information that can be transmitted per second by a channel. If a channel can transmit a maximum of K pulses per second, then the channel capacity C is given by

$$C = {}^{K}\!/_{2} \log_2\left(1 + {}^{S}\!/_{N}\right) bits/second$$

- A system of bandwidth nf_m Hz can transmit $2nf_m$ independent pulses per second. It is concluded that a system with bandwidth B Hz can transmit a maximum of 2B pulses per second. Replacing K with 2B, we eventually get

$$C = B \log_2\left(1 + {}^{S}\!/_{N}\right) bits/second$$

- The bandwidth and the signal power can be exchanged for one another.

The Source Coding:

- **Definition:** A conversion of the output of a discrete memory less source (DMS) into a sequence of binary symbols i.e. binary code word, is called **Source Coding**.
- The device that performs this conversion is called the **Source Encoder**.
- **Objective of Source Coding:** An objective of source coding is to minimize the average bit rate required for representation of the source by reducing the redundancy of the information source.

Few Terms Related to Source Coding Process:

I. Code word Length:

- Let X be a DMS with finite entropy H (X) and an alphabet $\{x_1 \dots \dots \ldots x_m\}$ with corresponding probabilities of occurrence P(x_i) (i = 1, …. , m). Let the binary code word assigned to symbol x_i by the encoder have length n_i, measured in bits. The length of the code word is the number of binary digits in the code word.

II. Average Code word Length:

- The average code word length L, per source symbol is given by

$$\boldsymbol{L = \int_{i=1}^{m} P(x_i) n_i}$$

- The parameter L represents the average number of bits per source symbol used in the source coding process.

III. Code Efficiency:

- The *code efficiency* η is defined as

$$\boldsymbol{\eta} = {}^{\boldsymbol{L_{min}}}/_{\boldsymbol{L}}$$

- where L_{min} is the minimum value of L. As η approaches unity, the code is said to be efficient.

IV. Code Redundancy:

- The *code redundancy* γ is defined as

$$\boldsymbol{\gamma} = \mathbf{1} - \eta$$

The Source Coding Theorem:

- The source coding theorem states that for a DMS X, with entropy H (X), the average code word length L per symbol is bounded as L ≥ H (X).
- And further, L can be made as close to H (X) as desired for some suitable chosen code.
- Thus, with

$$L_{min} = H\ (X)$$

- The code efficiency can be rewritten as

$$\eta = {}^{H(X)}/_{L}$$

Classification of Code:

- Fixed – Length Codes
- Variable – Length Codes
- Distinct Codes
- Prefix – Free Codes
- Uniquely Decodable Codes
- Instantaneous Codes
- Optimal Codes

x_i	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
x_1	00	00	0	0	0	1
x_2	01	01	1	10	01	01
x_3	00	10	00	110	011	001
x_4	11	11	11	111	0111	0001

- **Fixed – Length Codes:**

A fixed – length code is one whose code word length is fixed. Code 1 and Code 2 of above table are fixed – length code words with length 2.

- **Variable – Length Codes:**

A variable – length code is one whose code word length is not fixed. All codes of above table except Code 1 and Code 2 are variable – length codes.

- **Distinct Codes:**

A code is distinct if each code word is distinguishable from each other. All codes of above table except Code 1 are distinct codes.

- **Prefix – Free Codes:**

A code in which no code word can be formed by adding code symbols to another code word is called a prefix- free code. In a prefix – free code, no code word is prefix of another. Codes 2, 4 and 6 of above table are prefix – free codes.

- **Uniquely Decodable Codes:**

A distinct code is uniquely decodable if the original source sequence can be reconstructed perfectly from the encoded binary sequence. A sufficient condition to ensure that a code is uniquely decodable is that no code word is a prefix of another. Thus the prefix – free codes 2, 4 and 6 are uniquely decodable codes. Prefix – free condition is not a necessary condition for uniquely decidability. Code 5 albeit does not satisfy the prefix – free condition and yet it is a uniquely decodable code since the bit 0 indicates the beginning of each code word of the code.

- **Instantaneous Codes:**

A uniquely decodable code is called an instantaneous code if the end of any code word is recognizable without examining subsequent code symbols. The instantaneous codes have the property previously mentioned that no code word is a prefix of another code word. Prefix – free codes are sometimes known as instantaneous codes.

- **Optimal Codes:**

A code is said to be optimal if it is instantaneous and has the minimum average L for a given source with a given probability assignment for the source symbols.

Kraft Inequality:

- Let X be a DMS with alphabet $\{x_i\}(i = 1,2, \ldots , m)$. Assume that the length of the assigned binary code word corresponding to x_i is n_i.
- A necessary and sufficient condition for the existence of an instantaneous binary code is

$$K = \sum_{i=1}^{m} 2^{-n_i} \leq 1$$

- This is known as the **Kraft Inequality**.
- It may be noted that Kraft inequality assures us of the existence of an instantaneously decodable code with code word lengths that satisfy the inequality.
- But it does not show us how to obtain those code words, nor does it say any code satisfies the inequality is automatically uniquely decodable.

Entropy Coding:

- The design of a variable – length code such that its average code word length approaches the entropy of DMS is often referred to as **Entropy Coding**.

- There are basically two types of entropy coding, viz.

1. **Shannon – Fano Coding**
2. **Huffman Coding**

Shannon – Fano Coding:

- An efficient code can be obtained by the following simple procedure, known as **Shannon – Fano algorithm.**

1. List the source symbols in order of decreasing probability.
2. Partition the set into two sets that are as close to equiprobables as possible and assign 0 to the upper set and 1 to the lower set.
3. Continue this process, each time partitioning the sets with as nearly equal probabilities as possible until further partitioning is not possible.
4. Assign code word by appending the 0s and 1s from left to right.

Shannon – Fano Coding - Example

- Let there be six (6) source symbols having probabilities as $x_1 = 0.30$, $x_2 = 0.25$, $x_3 = 0.20$, $x_4 = 0.12$, $x_5 = 0.08$ $x_6 = 0.05$. Obtain the Shannon – Fano Coding for the given source symbols.

  - Shannon Fano Code words

x_i	P (x_i)	Step 1	Step 2	Step 3	Step 4	Code
x_1	0.30	0	0			00
x_2	0.25	0	1			01
x_3	0.20	1	0			10
x_4	0.12	1	1	0		110
x_5	0.08	1	1	1	0	1110
x_6	0.05	1	1	1	1	1111

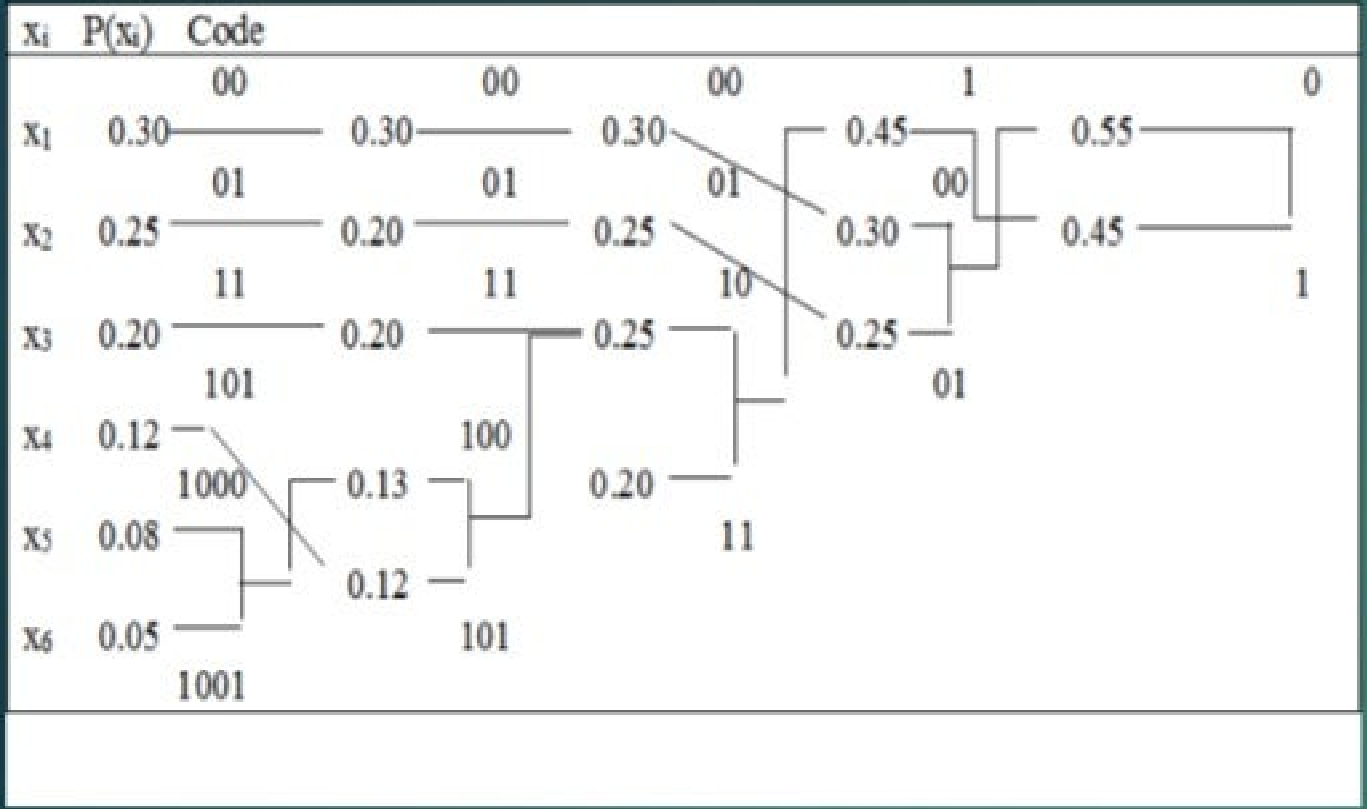
- **H (X) = 2.36 b/symbol**
- **L = 2.38 b/symbol**
- **η = H (X)/ L = 0.99**

Huffman Coding:

- Huffman coding results in an optimal code. It is the code that has the highest efficiency.
- The Huffman coding procedure is as follows:

1. List the source symbols in order of decreasing probability.
2. Combine the probabilities of the two symbols having the lowest probabilities and reorder the resultant probabilities, this step is called reduction 1. The same procedure is repeated until there are two ordered probabilities remaining.
3. Start encoding with the last reduction, which consists of exactly two ordered probabilities. Assign 0 as the first digit in the code word for all the source symbols associated with the first probability; assign 1 to the second probability.
4. Now go back and assign 0 and 1 to the second digit for the two probabilities that were combined in the previous reduction step, retaining all the source symbols associated with the first probability; assign 1 to the second probability.
5. Keep regressing this way until the first column is reached.
6. The code word is obtained tracing back from right to left.

Huffman Encoding - Example



Source Sample x_i	P (x_i)	Codeword
x_1	0.30	00
x_2	0.25	01
x_3	0.20	11
x_4	0.12	101
x_5	0.08	1000
x_6	0.05	1001

H (X) = 2.36 b/symbol

L = 2.38 b/symbol

η = H (X)/ L = 0.99

Redundancy:

- Redundancy in information theory refers to the reduction in information content of a message from its maximum value.
- For example, consider English having 26 alphabets. Assuming all alphabets are equally likely to occur, P (x_i) = 1/26. For all the 26 letters, the information contained is therefore

$$\log_2 26 = 4.7\ bits/letter$$

- Assuming that each letter to occur with equal probability is not correct, if we assume that some letters are more likely to occur than others, it actually reduces the information content in English from its maximum value of 4.7 bits/symbol.
- We define relative entropy on the ratio of H (Y/X) to H (X) which gives the maximum compression value and Redundancy is then expressed as

$$Redundancy = -\ {}^{H(Y/X)}\!\big/_{H(X)}$$

Entropy Relations for a continuous Channel:

- In a continuous channel, an information source produces a continuous signal x (t).
- The signal offered to such a channel of an ensemble of waveforms generated by some random ergodic processes if the channel is subject to AWGN.
- If the continuous signal x (t) has a finite bandwidth, it can be as well described by a continuous random variable X having a PDF (Probability Density Function) f_x (x).

- The entropy of a continuous source X (t)is given by

$$H\,(X) = -\int_{-\infty}^{\infty} f_x(x) \log_2 f_x(x) dx\ bits/symbol$$

- This relation is also known as **Differential Entropy of X**.

- Similarly, the average mutual information for a continuous channel is given by

$$I(X;Y) = H(X) - H\left(\frac{X}{Y}\right) = H(Y) - H(Y/X)$$

- Where

$$H(Y) = -\int_{-\infty}^{\infty} f_Y(y)\log_2 f_Y(y)dy$$

$$H\left(\frac{X}{Y}\right) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2 f_X\left({}^{x}/_{y}\right)dxdy$$

$$H\left(\frac{Y}{X}\right) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2 f_Y\left({}^{y}/_{x}\right)dxdy$$

Analytical Proof of Shannon – Hartley Law:

- Let x, A and y represent the samples of x (t), A(t) and y (t).
- Assume that the channel is band limited to B Hz. x (t) and A (t) are also limited to B Hz and hence y (t) is also bandlimited to B Hz.
- Hence all signals can be specified by taking samples at Nyquist rate of 2B samples/sec.
- We know

$$I(X;Y) = H(Y) - \ H(Y/X)$$

$$H\left(\frac{Y}{X}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x,y)\log {}^{1}/_{P(y/x)}\, dxdy$$

$$H\left(\frac{Y}{X}\right) = \int_{-\infty}^{\infty} P(x)dx \int_{-\infty}^{\infty} P(y/x)\log 1/P(y/x)\, dy$$

- We know that y = x + n.
- For a given x, y = n + a [a = constant of x]
- Hence, the distribution of y for which x has a given value is identical to that of ‘n’ except for a translation of x.

- If P_N (n) represents the probability density function of ‘n’, then
- P (y/x) = P_n (y - x)
- Therefore

$$\int_{-\infty}^{\infty} P(y/x)\log 1/P(y/x)\, dy = \int_{-\infty}^{\infty} P_n(y-x)\log 1/P_n(y-x)\, dy$$

- Let y – x = z; $dy = dz$
- Hence, the above equation becomes

$$\int_{-\infty}^{\infty} P_n(z)\log 1/P_n(z)\, dz = H(n)$$

- H (n) is the entropy of the noise sample.
- If the mean square value of the signal x (t) is S and that of noise n (t) is N, then the mean square value of the output is given by

$$\overline{y^2} = S + N$$

- H (n) is the entropy of the noise sample.
- If the mean square value of the signal x (t) is S and that of noise n (t) is N, then the mean square value of the output is given by $\overline{y^2} = S + N$
- Channel capacity C = max [I (X;Y)] = max [H(Y) – H(Y/X)] = max [H(Y) – H(n)]
- H(Y) is the maximum when y is Gaussian and the maximum value of H(Y) is f_m.log ($2\Pi e\sigma^2$); whereas f_m is the band limited frequency, σ^2 is the mean square value where σ^2 = S + N.
- Therefore,

$$H(y)_{max} = B\log 2\pi e\,(S+N)$$

- $\max[H(n)] = B\log 2\pi e(N)$
- Therefore

$$C = B\log[2\pi e(S+N)] - B\log(2\pi eN)$$

$$\text{Or } C = B\log[{}^{2\pi e(S+N)}/_{2\pi eN}]$$

- Hence

$$C = B\log[1 + {}^{S}/_{N}]$$

Rate Distortion Theory:

- By source coding theorem for a discrete memoryless source, according to which the average code – word length must be at least as large as the source entropy for perfect coding (i.e. perfect representation of the source).
- There are constraints that force the coding to be imperfect, thereby resulting in unavoidable distortion.
- For example, the source may have a continuous amplitude as in case of speech and the requirement is to quantize the amplitude of each sample generated by the source to permit its representation by a code word of finite length as in PCM.
- In such a case, the problem is referred to as Source coding with a fidelity criterion and the branch of information theory that deals with it is called **Rate Distortion Theory**.

- RDT finds applications in two types of situations:

1. Source coding where permitted coding alphabet can not exactly represent the information source, in which case we are forced to do ‘lossy data compression’.
2. Information transmission at a rate greater than the channel capacity.

Rate Distortion Function:

- Consider DMS defined by a M – ary alphabet X: {x_i| i = 1, 2, …., M} which consists of a set of statistically independent symbols together with the associated symbol probabilities {P_i | i = 1, 2, .. , M}. Let R be the average code rate in bits per code word.
- The representation code words are taken from another alphabet Y: {y_j| j = 1, 2, … N}.
- This source coding theorem states that this second alphabet provides a perfect representation of the source provided that R > H, where H is the source entropy.
- But if we are forced to have R < H, then there is an unavoidable distortion and therefore loss of information.
- Let P (x_i, y_j) denote the joint probability of occurrence of source symbol x_i and representation symbol y_j.

- From probability theory, we have
- P (x_i, y_j) = P (y_j/x_i) P (x_i) where P (y_j/x_i) is a transition probability.
- Let d (x_i, y_j) denote a measure of the cost incurred in representing the source symbol x_i by the symbol y_j;
- the quantity d (x_i, y_j) is referred to as a single – letter distortion measure.
- The statistical average of d(x_i, y_j) over all possible source symbols and representation symbols is given by

$$\overline{d} = \int_{i=1}^{M} \int_{j=1}^{N} P(x_i)\, P\left({}^{y_j}/_{x_i} \right) d(x_i y_j)$$

Average distortion $\overline{d}$ is a non – negative continuous function of the transition probabilities P (y_j / x_i) those are determined by the source encoder – decoder pair.

- A conditional probability assignment P (y$_j$ / x$_i$) is said to be D – admissible if and only the average distortion $\overline{d}$ is less than or equal to some acceptable value D. the set of all D – admissible conditional probability assignments is denoted by

$$P_D = \left\{ \mathrm{P}\left({}_{\mathrm{y_j|}} \mathrm{x_i} \right) \right\} : \overline{d} \leq D$$

- For each set of transition probabilities, we have a mutual information

$$I(X^{;} Y) = \int_{=1}^{M} \int_{=1}^{N} p(^{x_i}) p(y_j/x_i) \log(\overset{p(y_j/x_i)}{\quad} /p(y_j))$$

- "**<u>A Rate Distortion Function R (D)</u>** is defined as the smallest coding rate possible for which the average distortion is guaranteed not to exceed D".
- Let P_D denote the set to which the conditional probability P (y_j / x_i) belongs for prescribed D. then, for a fixed D, we write

$$R(D) = \min_{P(y_j / x_i) \in P_D} I(X;Y)$$

- Subject to the constraint

$$\sum_{j=1}^{N} P\left(\frac{y_j}{x_i}\right) = 1 \; for \; i = 1, 2, \dots., M$$

- The rate distortion function R (D) is measured in units of bits if the base – 2 logarithm is used.
- We expect the distortion D to decrease as R (D) is increased.
- Tolerating a large distortion D permits the use of a smaller rate for coding and/or transmission of information.

Gaussian Source:

- Consider a discrete time memoryless Gaussian source with zero mean and variance σ^2. Let x denote the value of a sample generated by such a source.
- Let y denote a quantized version of x that permits a finite representation of it.
- The squared error distortion $d(x, y) = (x-y)^2$ provides a distortion measure that is widely used for continuous alphabets.
- The rate distortion function for the Gaussian source with squared error distortion, is given by

  - $R\ (D) = \frac{1}{2} \log(\sigma^2/_D);\ 0 \leq D \leq \sigma^2$ otherwise 0 for $D > \sigma^2$

- In this case, $R\ (D) \rightarrow \infty$ as $D \rightarrow 0$
- $R\ (D) \rightarrow 0$ for $D = \sigma^2$

Data Compression:

- RDT leads to data compression that involves a purposeful or unavoidable reduction in the information content of data from a continuous or discrete source.
- Data compressor is a device that supplies a code with the least number of symbols for the representation of the source output subject to permissible or acceptable distortion.

Probability Theory:

$$0 \le {}^{N_n(A)}\!/_{n} \le 1$$

$$P(A) = \lim_{n\to\infty} {}^{N_n(A)}\!/_{n}$$

- **Axioms of Probability:**

1. P (S) = 1
2. 0 ≤ P (A) ≤ 1
3. P (A+ B) = P (A) + P (B)
4. $\{N_n\ (A+B)\}/n = N_n(A)/n + N_n(B)/n$

- **Property 1:**
- $P(A') = 1 - P(A)$
- $S = A + A'$
- $1 = P(A) + P(A')$

- **Property 2:**
- If M mutually exclusive events A_1, A_2, ……, $A_{m,}$ have the exhaustive property
- $A_1 + A_2 + \ldots\ldots + A_m = S$
- Then
- $P(A_1) + P(A_2) + \ldots.. P(A_m) = 1$

- **Property 3:**
- When events A and B are not mutually exclusive, then the probability of the union event ‘A or B’ equals
- P (A+B) = P (A) + P (B) – P (AB)
- Where P (AB) is the probability of the joint event ‘A and B’.
- P (AB) is called the ‘joint probability’ if

$$P\ (AB) = \lim_{n\to\infty} ({}^{N_n(AB)}/_{n})$$

Property 4:

- Let there involves a pair of events A and B.
- Let P (B/A) denote the probability of event B, given that event A has occurred and P (B/A) is called ‘conditional probability’ if

$$P\left(\frac{B}{A}\right) = {}^{P\ (AB)}/_{P(A)}$$

- We have N_n (AB)/ N_n (A) ≤ 1
- Therefore,

$$P(B?\,A) = \lim_{n\to\infty} ({}^{N_n(AB)}/_{N_n(A)})$$

References:

1. Digital Communications; Dr. Sanjay Sharma; Katson Books.
2. Communication Engineering; B P Lathi.