
📘 INFORMATION THEORY – Unit-1 NOTES

📌 1. Introduction to Information Theory

Information theory, developed by **Claude Shannon** in 1948, is the mathematical study of **data communication, compression, storage, and transmission**. It quantifies **information, entropy, redundancy, and channel capacity**, forming the basis of digital communication systems and coding theory.

Overview: What is Information Theory?

Key idea: The movements and transformations of information, just like those of a fluid, are constrained by mathematical and physical laws. These laws have deep connections with:

- **probability theory, statistics, and combinatorics**
- **thermodynamics (statistical physics)**
- **spectral analysis, Fourier (and other) transforms**
- **sampling theory, prediction, estimation theory**
- **electrical engineering (bandwidth; signal-to-noise ratio)**
- **complexity theory (minimal description length)**
- **signal processing, representation, compressibility**

As such, information theory addresses and answers the two fundamental questions of communication theory:

1. What is the ultimate data compression? (Answer: the entropy of the data, H, is its compression limit.)

2. What is the ultimate transmission rate of communication? (Answer: the channel capacity, C, is its rate limit.

All communication schemes lie in between these two limits on the compressibility of data and the capacity of a channel. Information theory can suggest means to achieve these theoretical limits. But the subject also extends far beyond communication theory.

Important questions… to which Information Theory offers answers:

• How should information be measured?

• How much additional information is gained by some reduction in uncertainty?

• How do the priori probabilities of possible messages determine the informativeness of receiving them?

• What is the information content of a random variable?

• How does the noise level in a communication channel limit its capacity to transmit information?

• How does the bandwidth (in cycles/second) of a communication channel limit its capacity to transmit information?

• By what formalism should prior knowledge be combined with incoming data to draw formally justifiable inferences from both?

• How much information is contained in a strand of DNA?

• How much information is there in the firing pattern of a neurone.

Historical origins and important contributions:

• **Ludwig BOLTZMANN (1844-1906),** physicist, showed in 1877 that thermodynamic entropy (defined as the energy of a statistical ensemble [such as a gas] divided by its temperature: ergs/degree) is related to the statistical distribution of molecular configurations, with increasing entropy corresponding to increasing randomness. He made this relationship precise with his famous formula S = k logW where S defines entropy, W is the total number of possible molecular configurations, and k is the constant which bears Boltzmann's name: k =1.38 x 10^{-16} ergs per degree centigrade. (The above formula appears as an epitaph on Boltzmann's tombstone.) This is equivalent to the definition of the information ("negentropy") in an ensemble, all of whose possible states are equiprobable, but with a minus sign in front (and when the logarithm is base 2, k=1.) The deep connections between Information Theory and that branch of physics concerned with thermodynamics and statistical mechanics, hinge upon Boltzmann's work.

- Leo SZILARD (1898-1964) in 1929 identified entropy with information. He formulated key information-theoretic concepts to solve the thermodynamic paradox known as "Maxwell's demon" (a thought-experiment about gas molecules in a partitioned box) by showing that the amount of information required by the demon about the positions and velocities of the molecules was equal (negatively) to the demon's entropy increment.
- James Clerk MAXWELL (1831-1879) originated the paradox called "Maxwell's Demon" which greatly influenced Boltzmann and which led to the watershed insight for information theory contributed by Szilard. At Cambridge, Maxwell founded the Cavendish Laboratory which became the original Department of Physics.
- R V HARTLEY in 1928 founded communication theory with his paper *Transmission of Information*. He proposed that a signal (or a communication channel) having bandwidth Ω over a duration T has a limited number of degrees-of-freedom, namely $2\Omega T$, and therefore it can communicate at most this quantity of information. He also defined the information content of an equiprobable ensemble of N possible states as equal to $\log_2 N$.
- Norbert WIENER (1894-1964) unified information theory and Fourier analysis by deriving a series of relationships between the two. He invented "white noise analysis" of non-linear systems, and made the definitive contribution to modeling and describing the information content of stochastic processes known as *Time Series*.

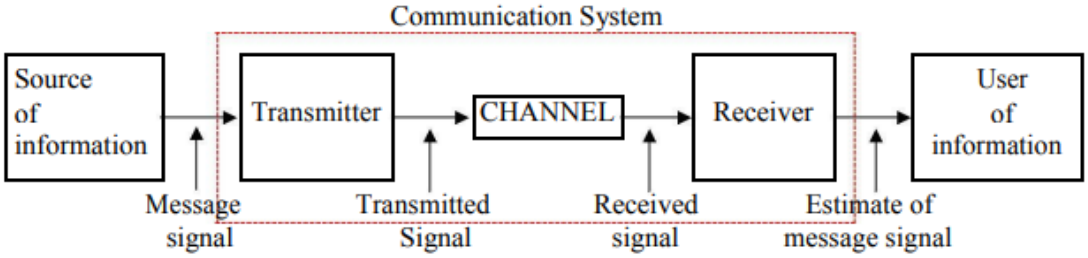
- Dennis GABOR (1900-1979) crystallized Hartley's insight by formulating a general *Uncertainty Principle* for information, expressing the trade-off for resolution between bandwidth and time. (Signals that are well specified in frequency content must be poorly localized in time, and those that are well localized in time must be poorly specified in frequency content.) He formalized the "Information Diagram" to describe this fundamental trade-off, and derived the continuous family of functions which optimize (minimize) the conjoint uncertainty relation. In 1974 Gabor won the Nobel Prize in Physics for his work in Fourier optics, including the invention of holography.
- Claude SHANNON (together with Warren WEAVER) in 1949 wrote the definitive, classic, work in information theory: *Mathematical Theory of Communication.* Divided into separate treatments for continuous-time and discrete-time signals, systems, and channels, this book laid out all of the key concepts and relationships that define the field today. In particular, he proved the famous Source Coding Theorem and the Noisy Channel Coding Theorem, plus many other related results about channel capacity.
- S KULLBACK and R A LEIBLER (1951) defined *relative entropy* (also called *information for discrimination,* or *K-L Distance.*)
- E T JAYNES (since 1957) developed *maximum entropy* methods for inference, hypothesis-testing, and decision-making, based on the physics of statistical mechanics. Others have inquired whether these principles impose fundamental physical limits to computation itself.
- A N KOLMOGOROV in 1965 proposed that the *complexity* of a string of data can be defined by the length of the shortest binary program for computing the string. Thus the complexity of data is its *minimal description length,* and this specifies the ultimate compressibility of data. The "Kolmogorov complexity" K of a string is approximately equal to its Shannon entropy H, thereby unifying the theory of descriptive complexity and information theory.

1.1 Introduction:

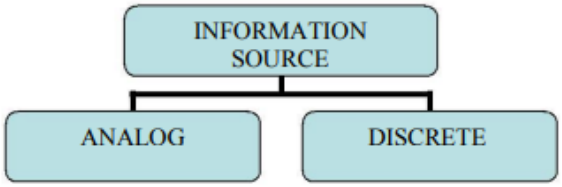
- **Communication**

Communication involves explicitly the transmission of information from one point to another, through a succession of processes.

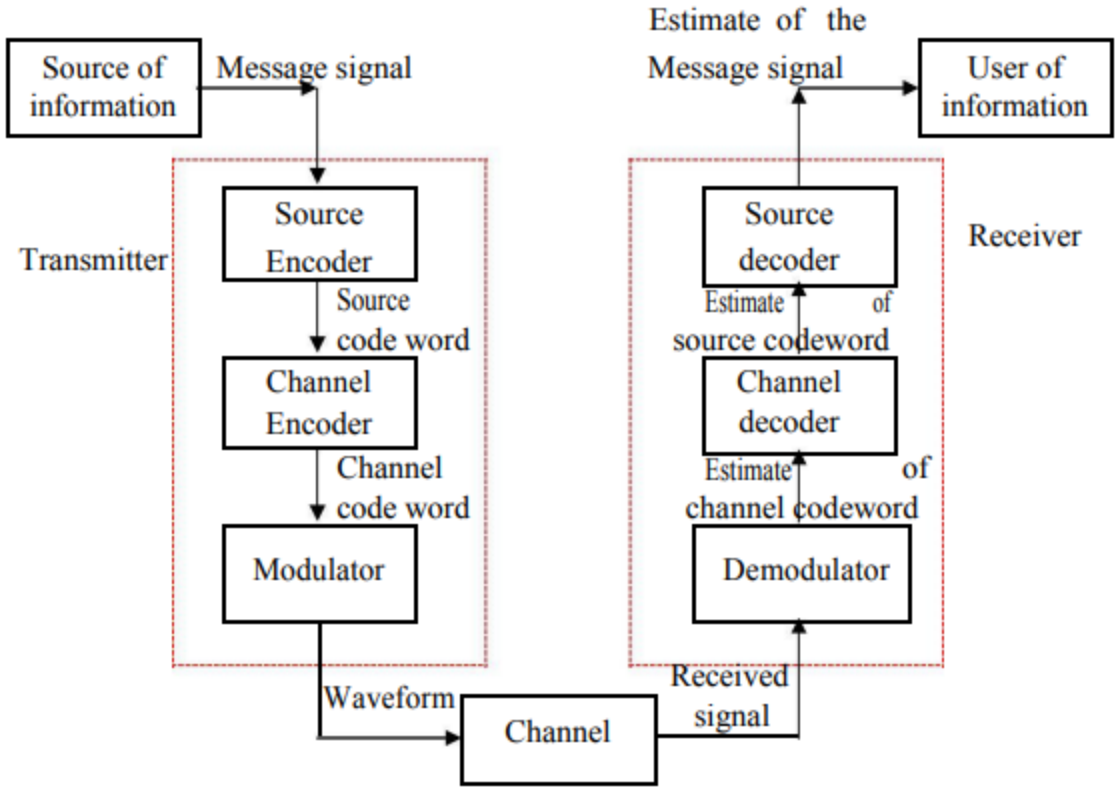
- **Basic elements to every communication system**
  - Transmitter
  - Channel and
  - Receiver



- **Information sources are classified as:**



Digital Communication System:



- **Source definition**

  Analog : Emit a continuous – amplitude, continuous – time electrical wave from. Discrete : Emit a sequence of letters of symbols.

  The output of a discrete information source is a string or sequence of symbols.

1.2 Measure the information:

To measure the information content of a message quantitatively, we are required to arrive at an intuitive concept of the amount of information.

Consider the following examples:

A trip to Mercara (Coorg) in the winter time during evening hours,

1. It is a cold day
2. It is a cloudy day
3. Possible snow flurries

2. Mathematical Foundations; Probability Rules; Bayes' Theorem

What are random variables? What is probability?

Random variables are variables that take on values determined by probability distributions. They may be discrete or continuous, in either their domain or their range. For example, a stream of ASCII encoded text characters in a transmitted message is a discrete random variable, with a known probability distribution for any given natural language. An analog speech signal represented by a voltage or sound pressure waveform as a function of time (perhaps with added noise), is a continuous random variable having a continuous probability density function.

Most of Information Theory involves probability distributions of random variables, and conjoint or conditional probabilities defined over ensembles of random variables. Indeed, the information content of a symbol or event is defined by its (im)probability. Classically, there are two different points of view about what probability actually means:

- *relative frequency:* sample the random variable a great many times and tally up the fraction of times that each of its different possible values occurs, to arrive at the probability of each.
- *degree-of-belief:* probability is the plausibility of a proposition or the likelihood that a particular state (or value of a random variable) might occur, even if its outcome can only be decided once (e.g. the outcome of a particular horse-race).

The first view, the "frequentist" or operationalist view, is the one that predominates in statistics and in information theory. However, by no means does it capture the full meaning of probability. For example, the proposition that `"The moon is made of green cheese"` is one which surely has a probability that we should be able to attach to it. We could assess its probability by degree-of-belief calculations which

combine our prior knowledge about physics, geology, and dairy products. Yet the "frequentist" definition of probability could only assign a probability to this proposition by performing (say) a large number of repeated trips to the moon, and tallying up the fraction of trips on which the moon turned out to be a dairy product....

In either case, it seems sensible that the less probable an event is, the more information is gained by noting its occurrence. (Surely discovering that the moon IS made of green cheese would be more "informative" than merely learning that it is made only of earth-like rocks.)

Probability Rules

Most of probability theory was laid down by theologians: Blaise PASCAL (1623-1662) who gave it the axiomatization that we accept today; and Thomas BAYES (1702-1761) who expressed one of its most important and widely-applied propositions relating conditional probabilities.

Probability Theory rests upon two rules:

Product Rule:

$$\begin{aligned} p(A,B) &= \text{"joint probability of } \textit{both } A \textit{ and } B\text{"} \\ &= p(A|B)p(B) \end{aligned}$$

$$\begin{aligned} &\text{or equivalently,} \\ &= p(B|A)p(A) \end{aligned}$$

Clearly, in case A and B are *independent* events, they are not conditionalized on each other and so

$$\begin{aligned} p(A|B) &= p(A) \\ \text{and } p(B|A) &= p(B), \end{aligned}$$

in which case their joint probability is simply $p(A,B) = p(A)p(B)$.

Sum Rule:

If event A is conditionalized on a number of other events B, then the total probability of A is the sum of its joint probabilities with all B:

$$p(A) = \sum_B p(A, B) = \sum_B p(A|B)p(B)$$

From the Product Rule and the symmetry that $p(A, B) = p(B, A)$, it is clear that $p(A|B)p(B) = p(B|A)p(A)$. Bayes' Theorem then follows:

Bayes' Theorem:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

The importance of Bayes' Rule is that it allows us to reverse the conditionalizing of events, and to compute $p(B|A)$ from knowledge of $p(A|B)$, $p(A)$, and $p(B)$. Often these are expressed as *prior* and *posterior* probabilities, or as the conditionalizing of hypotheses upon data.

Basic Terminologies

Term	Description
Information	Reduction in uncertainty due to reception of a message.
Source	The system which generates the data (e.g., English text, binary stream).
Entropy (H)	Average amount of information per symbol.
Redundancy	Excess bits used for error detection/correction.
Channel	Medium through which information is transmitted.
Noise	Any unwanted disturbance on the channel.

Marginal probability: From the Sum Rule, we can see that the probability of X taking on a particular value $x = a_i$ is the sum of the joint probabilities of this outcome for X and all possible outcomes for Y:

$$p(x = a_i) = \sum_y p(x = a_i, y)$$

We can simplify this notation to: $p(x) = \sum_y p(x, y)$

and similarly: $p(y) = \sum_x p(x, y)$

Conditional probability: From the Product Rule, we can easily see that the conditional probability that $x = a_i$, given that $y = b_j$, is:

$$p(x = a_i | y = b_j) = \frac{p(x = a_i, y = b_j)}{p(y = b_j)}$$

We can simplify this notation to: $p(x|y) = \frac{p(x, y)}{p(y)}$

and similarly: $p(y|x) = \frac{p(x, y)}{p(x)}$

It is now possible to define various entropy measures for joint ensembles:

3. Entropies Defined, and Why They are Measures of Information

The information content I of a single event or message is defined as the base-2 logarithm of its probability p:

$$I = \log_2 p \tag{1}$$

and its *entropy* H is considered the negative of this. Entropy can be regarded intuitively as "uncertainty," or "disorder." To gain information is to lose uncertainty by the same amount, so I and H differ only in sign (if at all): $H = -I$. Entropy and information have units of *bits.*

Note that I as defined in Eqt (1) is never positive: it ranges between 0 and $-\infty$ as p varies from 1 to 0. However, sometimes the sign is dropped, and I is considered the same thing as H (as we'll do later too).

No information is gained (no uncertainty is lost) by the appearance of an event or the receipt of a message that was completely certain anyway ($p = 1$, so $I = 0$). Intuitively, the more improbable an event is, the more informative it is; and so the monotonic behaviour of Eqt (1) seems appropriate. But why the logarithm?

The logarithmic measure is justified by the desire for information to be additive. We want the algebra of our measures to reflect the Rules of Probability. When independent packets of information arrive, we would like to say that the total information received is the sum of the individual pieces. But the probabilities of independent events multiply to give their combined probabilities, and so we must take logarithms in order for the joint probability of independent events or messages to contribute additively to the information gained.

This principle can also be understood in terms of the combinatorics of state spaces. Suppose we have two independent problems, one with n

possible solutions (or states) each having probability p_n, and the other with m possible solutions (or states) each having probability p_m. Then the number of combined states is mn, and each of these has probability $p_m p_n$. We would like to say that the information gained by specifying the solution to *both* problems is the *sum* of that gained from each one. This desired property is achieved:

$$I_{mn} = \log_2(p_m p_n) = \log_2 p_m + \log_2 p_n = I_m + I_n \qquad (2)$$

A Note on Logarithms:

In information theory we often wish to compute the base-2 logarithms of quantities, but most calculators (and tools like xcalc) only offer Napierian (base 2.718...) and decimal (base 10) logarithms. So the following conversions are useful:

$$\log_2 X = 1.443 \log_e X = 3.322 \log_{10} X$$

Henceforward we will omit the subscript; base-2 is always presumed.

📌 Measure of Information

If an event `E` has a probability `P(E)`, the amount of information `I(E)` associated with it is:

$$I(E) = -\log_b P(E)$$

Where `b` is the base of the logarithm (usually base 2 for binary systems, giving result in bits).

📌 Entropy (Average Information Content)

If a source emits `n` symbols $s_1, s_2, \ldots, s_n$ with respective probabilities $p_1, p_2, \ldots, p_n$, the **entropy** is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{(in bits)}$$

- **Maximum Entropy**: When all symbols are equally probable.
- **Minimum Entropy**: When one symbol is certain (probability = 1).

📌 5. Joint Entropy

For two random variables X and Y, the **joint entropy** is:

$$H(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y)$$

📌 6. Conditional Entropy

The uncertainty remaining in X given that Y is known:

$$H(X|Y) = -\sum_{x,y} p(x, y) \log_2 p(x|y)$$

📌 7. Mutual Information

Measures the amount of information that X and Y share:

$$I(X;Y) = H(X) - H(X|Y)$$

It quantifies the reduction in uncertainty of one variable due to knowledge of another.

Mutual Information between X and Y

The *mutual information* between two random variables measures the amount of information that one conveys about the other. Equivalently, it measures the average reduction in uncertainty about X that results from learning about Y. It is defined:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{12}$$

Clearly X says as much about Y as Y says about X. Note that in case X and Y are independent random variables, then the numerator inside the logarithm equals the denominator. Then the log term vanishes, and the mutual information equals zero, as one should expect.

Non-negativity: mutual information is always ≥ 0. In the event that the two random variables are perfectly correlated, then their mutual information is the entropy of either one alone. (Another way to say this is: $I(X;X) = H(X)$: the mutual information of a random variable with itself is just its entropy. For this reason, the entropy $H(X)$ of a random variable X is sometimes referred to as its *self-information*.)

📌 8. Channel Capacity

The **maximum rate** at which information can be transmitted over a noisy channel:

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

- For a **Binary Symmetric Channel (BSC)** with crossover probability `p`:

$$C = 1 - H(p)$$

📌 9. Source Coding Theorem

According to this theorem, "A message from a source with entropy H can be compressed to H bits/symbol **without loss**, on average."

📌 10. Noisy Channel Coding Theorem

It is possible to **transmit data with arbitrarily low error probability** over a noisy channel if the **transmission rate R ≤ channel capacity C**.

📌 11. Redundancy

Redundancy `R` is defined as:

$$R = 1 - \frac{H(X)}{\log_2 m}$$

Where `m` is the number of symbols in the alphabet.

Symbol:- In information theory and coding, a symbol is a discrete element or unit that represents information. It can be a letter, number, or any other predefined character within a specific alphabet or set. These symbols are used to construct messages or data sequences, which are then processed and transmitted.

📌 12. Applications of Information Theory

- **Data Compression** (e.g., Huffman, Shannon-Fano, Arithmetic coding)
- **Error Detection and Correction** (e.g., Hamming codes, CRC)
- **Cryptography**
- **Machine Learning** (e.g., Information Gain)
- **Image and Signal Processing**
- **Natural Language Processing**

📌 13. Common Coding Techniques

Technique	Purpose
Huffman Coding	Optimal prefix code for lossless compression.
Shannon-Fano Coding	Early method of entropy encoding.
Arithmetic Coding	Achieves higher compression rates than Huffman in some cases.
Run Length Encoding (RLE)	Compresses sequences of repeated values.
Lempel-Ziv-Welch (LZW)	Dictionary-based compression used in GIFs and TIFFs.

📌 **14. Information vs. Data**

Aspect	Information	Data
Meaning	Processed and meaningful	Raw and unprocessed
Value	High (used in decision-making)	Low (needs interpretation)

📌 **15. Summary-** Information theory provides the **foundation** for all digital communications, guiding how **data is encoded, compressed, transmitted, and decoded**. Its core principles like **entropy**, **mutual information**, and **channel capacity** are central to designing efficient and reliable communication systems.

Joint Entropy-

📌 **Definition:**

Joint Entropy is the measure of the total uncertainty associated with a pair of random variables X and Y taken together.

🔢 **Mathematical Expression:**

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y)$$

- Where $p(x,y)$ is the **joint probability** of the events $X = x$ and $Y = y$.
- The base of the logarithm determines the unit (base 2 for bits).

💡 **Interpretation:**

- Represents the **total information** required to describe the outcome of both X and Y together.
- If X and Y are independent:

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

Conditional Entropy (H(X|Y)) :-

Conditional Entropy is the **amount of uncertainty remaining in a random variable** X when the value of another variable Y is known.

🔢 **Mathematical Expression:**

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y)$$

🧠 **Intuition:**

- Measures **average uncertainty in** X given knowledge of Y.
- If X and Y are independent:

$$H(X|Y) = H(X)$$

- If X is completely determined by Y:

$$H(X|Y) = 0$$

Mutual Information (I(X; Y)) :-

📌 **Definition:**

Mutual Information measures the **amount of information** one random variable contains about another. It quantifies the **reduction in uncertainty** of one variable due to the knowledge of another.

🔢 **Mathematical Expression:**

$$I(X;Y) = H(X) - H(X|Y)$$

Alternate equivalent forms:

$$I(X;Y) = H(Y) - H(Y|X)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

✅ **Properties:**

- $I(X;Y) \geq 0$
- $I(X;Y) = 0$ if and only if X and Y are **independent**
- Symmetry: $I(X;Y) = I(Y;X)$

Applications in Information Theory

- Channel Capacity Analysis
- Feature Selection in Machine Learning
- Dependency Measurement in Data Mining
- Cryptography and Security
- Natural Language Processing

Information rate-

If the source is emitting symbols at a fixed rate of ''r_s ' symbols / sec, the average source information rate 'R' is defined as – R = r_s . H bits / sec

---

Topic: Markoff Statistical Model for Information Source, Entropy, and Information Rate

1. Introduction

The **Markoff statistical model** (also known as the **Markov model**) is widely used in information theory to model a source that emits symbols probabilistically, where the probability of each symbol depends on one or more previous symbols. Unlike a memoryless source, which emits each symbol independently, a **Markoff source** has memory and dependencies among symbols.

---

2. Markoff Statistical Model for Information Source

A **Markoff source** is characterized by the following:

Definition:

A source is said to be a **Markoff source of order *k*** if the probability of occurrence of a symbol depends only on the preceding *k* symbols.

For **first-order Markoff sources**, this simplifies to:

$$P(X_n|X_{n-1}, X_{n-2}, ..., X_1) = P(X_n|X_{n-1})$$

Key Elements:

- **Alphabet**: Set of possible symbols emitted (e.g., {A, B, C, ...}).
- **States**: Each state represents a symbol or group of symbols.
- **Transition Probabilities**:
  Pij=P(next symbol is j | current symbol is i)P_{ij} = P(\text{next symbol is } j \mid \text{current symbol is } i)Pij=P(next symbol is j|current symbol is i)

Transition Matrix:

A square matrix representing probabilities of transitioning from one symbol (state) to another.Example:

	A	B	C
A	0.2	0.5	0.3
B	0.1	0.6	0.3
C	0.3	0.3	0.4

3. Entropy of a Markoff Source

Definition: The **entropy** of a Markoff source measures the **average uncertainty** per symbol, taking into account dependencies between symbols. For a **first-order Markoff source**, the entropy HHH is defined as:

$$H = -\sum_i \pi_i \sum_j P_{ij} \log_2 P_{ij}$$

Where:

π_i: Stationary probability of state *i*

P_{ij}: Transition probability from state *i* to state *j*

Stationary Distribution:

The set of probabilities $\{\pi_i\}$ satisfying:

$$\pi_j = \sum_i \pi_i P_{ij}, \quad \sum_i \pi_i = 1$$

This represents the long-term average proportion of time the source spends in each state.

---

4. Information Rate of a Markoff Source

The **information rate** (also called **entropy rate**) represents the **average number of bits per symbol** produced by the source.

$$R = H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, ..., X_n)$$

For a **first-order Markoff source**, this simplifies to:

$$R = H = -\sum_i \pi_i \sum_j P_{ij} \log_2 P_{ij}$$

Thus, the **information rate is equal to the entropy** of the Markoff source when in stationary state.

---

5. Comparison: Memoryless vs. Markoff Source

Feature	Memoryless Source	Markoff Source
Dependency	No memory (independent symbols)	Depends on previous symbols
Entropy Formula	$H = -\sum p_i \log_2 p_i$	Requires stationary distribution
Information Rate	Constant per symbol	Depends on transition structure
Modeling Capability	Limited	Realistic for natural languages

---

6. Applications

- Text prediction and speech recognition
- Natural language processing
- Data compression algorithms (e.g., PPM, LZ78 variants)
- Hidden Markov Models (HMM) for pattern recognition

---

7. Conclusion

The **Markoff statistical model** allows for more accurate modeling of real-world sources by considering dependencies between symbols. Its entropy and information rate are essential measures in understanding the efficiency and capacity of communication systems.

---

8. References

1. Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. Wiley.
2. Viterbi, A. J., & Omura, J. K. (1979). *Principles of Digital Communication and Coding*. McGraw-Hill.
3. Sayood, K. (2017). *Introduction to Data Compression*. Morgan Kaufmann.

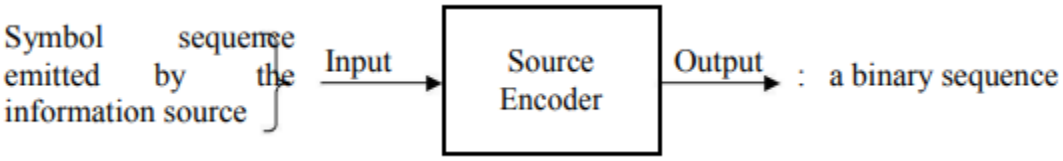
UNIT-2

Chapter: Source Coding and Communication Channels

---

1. Source Coding: Encoding of the Source Output

Source Coding refers to the process of representing the output of a source in binary form efficiently, i.e., reducing the redundancy while maintaining lossless information.



Where, G_N = –

- **If the encoder operates on blocks of 'N' symbols, t he bit rate of the encoder is given as**

Produces an average bit rate of G_N bits / symbol

$$\frac{1}{N}\sum_{i} p(m_i)\log p(m_i)$$

$p(m_i)$ = Probability of sequence **'m_i'** of **'N'** symbols from the source,

Sum is over all sequences 'm_i' containing **'N'** symbols.

G_N in a monotonic decreasing function of **N** and

$$\lim_{N\to\infty} G_N = H \text{ bits / symbol}$$

Performance measuring factor for the encoder

Coding efficiency: η_c

$$\text{Definition of } \eta c = \frac{\text{Source inf ormation rate}}{\text{Average output bit rate of the encoder}}$$

$$\eta_c = \frac{H(S)}{H_N}$$

Key Points:

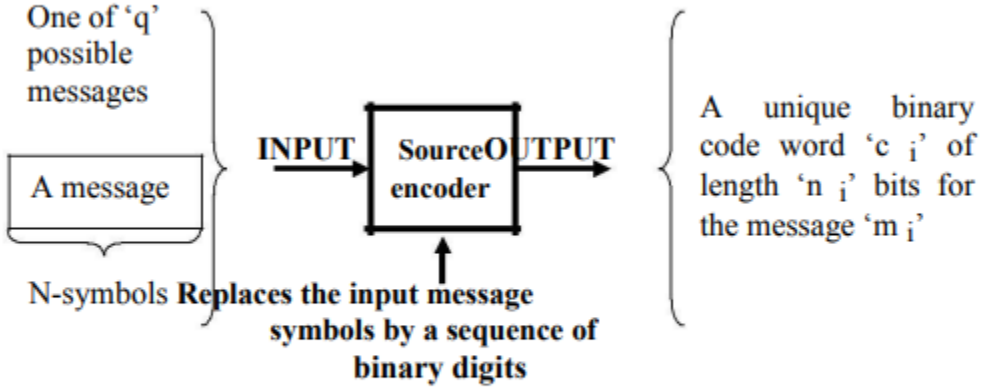
- The goal is to represent data with as few bits as possible.
- Achieved using variable-length codes depending on symbol probabilities.
- High-probability symbols → shorter codes; Low-probability symbols → longer codes.
- Examples: Shannon coding, Huffman coding, Arithmetic coding.

---

2. Shannon's Encoding Algorithm

- **Formulation of the design of the source encoder**

Can be formulated as follows:



'q' messages : $m_1, m_2, \ldots m_i, \ldots, m_q$

Probs. of messages : $p_1, p_2, \ldots p_i, \ldots, p_q$

n_i : an integer

Shannon's method encodes symbols based on their probabilities. The algorithm uses cumulative probabilities and binary fractions to generate prefix-free codes.

Steps:

1. Sort symbols in decreasing order of probability.
2. Calculate cumulative probability $F(x_i)$ for each symbol.
3. Code length:

$$l_i = \lceil -\log_2(p_i) \rceil$$

4. Code: Use binary expansion of $F(x_i) + \frac{p_i}{2}$, up to l_i bits.

- **Properties:**
- Simple and fast.
- Not always optimal, but close to entropy.
- Produces prefix codes.

🔶 3. Communication Channels

A **communication channel** is a medium through which information is transmitted from a sender (source) to a receiver (destination).

Components:

- Input (source symbols)
- Output (received symbols)
- Transition probabilities (channel behavior)

- **The schematic of a practical communication system is shown.**

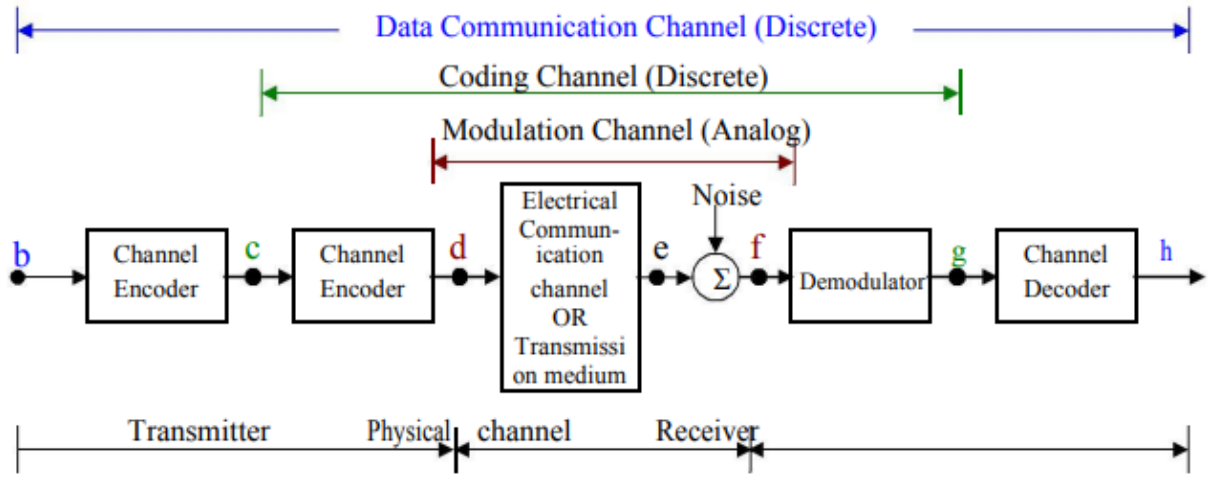


Fig.: BINARY COMMN. CHANNEL CHARACTERISATION

4. Discrete Communication Channels

These channels deal with **discrete** inputs and outputs (finite alphabets).

Example:

- Binary Symmetric Channel (BSC)

- Binary Erasure Channel (BEC)

Properties:

- Described using a transition probability matrix.
- Useful in digital communication systems.

---

5. Continuous Channels

These channels deal with **continuous-valued** signals (e.g., voltages, frequencies). a continuous channel as one whose input is a sample point from a continuous sample space and the output is a sample point belonging to either the same sample space or to a different sample space. Further we shall define a' zero memory continuous channel' as the one in which the channel output statistically depends on the corresponding channels without memory.

Example:

- Additive White Gaussian Noise (AWGN) channel.

Characterized by:

- Input/output signals being real-valued.
- Probabilistic models using probability density functions.

Entropy of continuous Signals: (Differential entropy): For the case of discrete messages, we have defined the entropy as,

$$H(S) = \sum_{k=1}^{q} p(s_k) \log \frac{1}{p(s_k)}$$

Chapter: Fundamental Limits on Performance

🔶 1. Source Coding Theorem (Shannon's First Theorem)

Statement:
For a discrete memoryless source with entropy H(X) , no lossless code can encode the source with average length less than H(X) bits/symbol.

$$H(X) \leq L < H(X) + 1$$

Implications:

- Entropy is the lower bound on average codeword length.
- It is possible to compress data close to its entropy using efficient coding schemes.

🔷 2. Huffman Coding

An optimal prefix coding algorithm that minimizes average codeword length.

Steps:

1. Create a priority queue of symbols by probability.
2. Combine two least-probable symbols into a node.
3. Repeat until a binary tree is formed.
4. Assign ‘0’ and ‘1’ to left/right edges.

Properties:

- Always produces optimal prefix codes.
- Greedy algorithm.
- Practical and widely used.

🔶

3. Discrete Memoryless Channels (DMC)

We have considered the discrete source, now we consider a channel through which we wish to pass symbols generated by such a source by some appropriate encoding mechanism; we also introduce the idea of noise into the system – that is we consider the channel to modify the input coding and possibly generate some modified version.

We should distinguish between systematic modification of the encoded symbols, i.e. distortion, and noise. Distortion is when an input code always results in the the same output code; this process can clearly be reversed. Noise on the other hand introduces the element of randomness into the resulting output code.

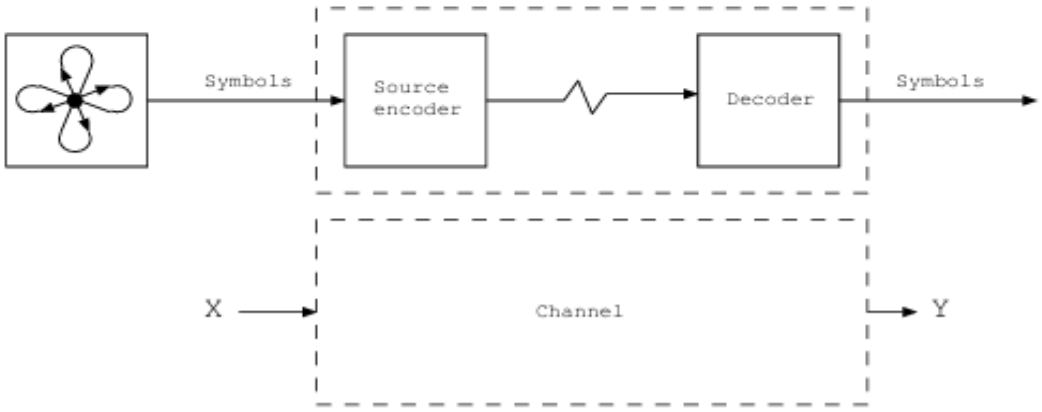


Figure- Coding and decoding of symbols for transfer over the channel

We consider an input alphabet $\mathcal{X} = \{x_1, \ldots, x_J\}$ and output alphabet $\mathcal{Y} = \{y_1, \ldots, y_K\}$ and random variables X and Y which range over these alphabets. Note that J and K need not be the same – for example we may have the binary input alphabet $\{0, 1\}$ and the output alphabet $\{0, 1, \perp\}$, where $\perp$ represents the decoder identifying some error. The discrete memoryless channel can then be represented as a set of transition probabilities:

$$p(y_k|x_j) = P(Y = y_k|X = x_j)$$

A **Discrete Memoryless Channel (DMC)** has:

- Discrete input/output alphabets.
- Memoryless: Output depends only on current input, not previous ones.

Represented by:

A transition probability matrix $P(Y|X)$

Examples:

- Binary Symmetric Channel (BSC)
- Z-Channel
- Erasure Channel

Binary Symmetric Channel:-

The binary symmetric channel has two input and output symbols (usually written $\{0, 1\}$) and a common probability, p, of "incorrect" decoding of an input at the output; this could be a simplistic model of a communications link, figure 20a.

However, to understand the averaging property of the error rate P_e described above, consider the figure 20b, where we have 10^6 symbols, of which the first has a probability of being received in error (of 0.1), and the remainder are always received perfectly. Then observing that most of the terms in the sum on the right of equation 46 are zero:

$$\begin{aligned} P_e &= p(y_1|x_0)p(x_0) \\ &= 0.1 \times 10^{-6} \end{aligned}$$

$$P_e = 10^{-7}$$

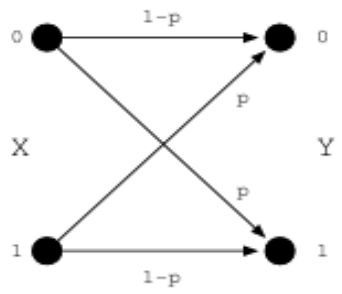


Figure- Binary Symmetric channel

Channel Capacity

We wish to define the capacity of a channel, using the model of a free input alphabet and dependent output alphabet (given by the channel matrix). We note that for a given channel, if we wish to maximize the mutual information, we must perform a maximization over all probability distributions for the input alphabet X. We define the *channel capacity*, denoted by C, as:

$$C = \max_{\{p(x_j)\}} I(\mathcal{X};\mathcal{Y}) \tag{51}$$

When we achieve this rate we describe the source as being matched to the channel.

Definition:
Maximum rate at which information can be reliably transmitted over a channel.

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

Where:

- C: Channel capacity (bits per use)
- $I(X;Y)$: Mutual information between input and output.

Capacity for Common Channels:

- **BSC (p):**

$$C = 1 - H(p) = 1 + p\log_2(p) + (1-p)\log_2(1-p)$$

- **AWGN Channel:**

$$C = \frac{1}{2}\log_2\left(1 + \frac{P}{N}\right) \quad \text{(bits per transmission)}$$

Channel Coding:-

To overcome the problem of noise in the system, we might consider adding redundancy during the encoding process to overcome possible errors. The examples that are used here are restricted to sources which would naturally be encoded in a noiseless environment as fixed size block codes – i.e. a source alphabet $\mathcal{X}$, which has 2^n equiprobable symbols; however, the discussion applies to more general sources and variable length coding schemes.

One particular aspect to be considered in real uses of channel coding is that many

---

sources which we are interested in encoding for transmissions have a significant amount of redundancy already. Consider sending a piece of syntactically correct and semantically meaningful English or computer program text through a channel which randomly corrupted on average 1 in 10 characters (such as might be introduced by transmission across a rather sickly Telex system). e.g.:

1. Bring reinforcements, we're going to advance
2. It's easy to recognise speech

Reconstruction from the following due to corruption of 1 in 10 characters would be comparatively straight forward:

1. Brizg reinforce ents, we're going to advance
2. It's easy mo recognise speech

However, while the redundancy of this source protects against such random character error, consider the error due to a human mis-hearing:

1. Bring three and fourpence, we're going to a dance.
2. It's easy to wreck a nice peach.

The coding needs to consider the error characteristics of the channel and decoder, and try to achieve a significant "distance" between plausible encoded messages.

Summary Table

Concept	Description
Source Coding	Compress data from source using binary codes
Shannon Encoding	Uses cumulative probability to assign binary codes
Communication Channel	Medium for sending symbols
Discrete Channels	Finite symbols, DMCs
Continuous Channels	Real-valued signals (e.g., AWGN)
Source Coding Theorem	$H(X) \leq L < H(X) + 1$
Huffman Coding	Optimal prefix code using greedy approach
DMC	Discrete input/output, no memory
Channel Capacity	Max reliable transmission rate over a channel

---

📚 References:

- Claude E. Shannon, “A Mathematical Theory of Communication”, 1948.
- Cover & Thomas, *Elements of Information Theory*.
- Simon Haykin, *Communication Systems*.