CHAPTER 1

INTRODUCTION

1.1 Overview of Information Theory

Information theory is a relatively new branch of mathematics that was made mathematically rigorous only in the 1940s. The term 'Information Theory' does not possess a unique definition. Broadly speaking, Information theory deals with the study of problems concerning any system. This includes information processing, information storage, information retrieval and decision making. In a narrow sense, Information theory studies all theoretical problems connected with the transmission of information over communication channels. This includes the study of uncertainty (information) measures and various practical and economical methods of coding information for transmission. Information Theory was developed by Claude E. Shannon to find fundamental limits on signal processing operations such as compressing data and on reliably storing and communicating data. Since its inception it has broadened to find applications in many other areas including statistical inference, natural language processing, cryptography, neurobiology, the evolution and function of molecular codes, model selection in ecology, thermal physics, quantum computing, plagiarism detection and other forms of data analysis.

Applications of fundamental topics of Information theory include lossless data compression (e.g. zip files), lossy data compression (e.g. mp3s and jpgs), and channel coding. The field is at the intersection of mathematics, statistics, computer science, physics, neurobiology, and electrical engineering. Its impact has been crucial to the success of the voyager missions to deep space, the invention of the compact disc, the feasibility of mobile phones, the development of the Internet, the study of linguistics and of human perception, the understanding of black holes, and numerous other fields. Important sub-fields of Information Theory are source coding, channel coding, algorithmic complexity

theory, algorithmic information theory, information-theoretic security, and measures of information. A key measure of information is entropy, which is usually expressed by the average number of bits needed to store or communicate one symbol in a message. Entropy quantifies the uncertainty involved in predicting the value of a random variable. For example, specifying the outcome of a fair coin flip (two equally likely outcomes) provides less information (lower entropy) than specifying the outcome from a roll of a die (six equally likely outcomes).

1.2 Basics Concepts and Definitions

1.2.1 Shannon Entropy

The concept of Shannon's entropy is the central role of Information Theory sometimes referred as measure of uncertainty. The entropy of a random variable is defined in terms of its probability distribution and can be a good measure of randomness or uncertainty.

In 1948, C.E. Shannon by profession an electrical engineer, was very much interested to communicate an information/ message across a noisy channel. Basically, he was interested to measure the loss of information supplied. He find that loss of information supplied = uncertainty removed. However, uncertainty is associated with every probability distribution viz. $P = p_1, p_2,..., p_n$, where $p_1, p_2,..., p_n$ are the probabilities of n outcomes. The uncertainty can, therefore, be measured by a function of $p_1, p_2,..., p_n$. Shannon used the axiomatic approach of Euclid in deriving his measure. He laid down some postulates for defining this measure of uncertainty $H(P)$ for a probability distribution P which are as follows:

(i) $H(P)$ should be a function of $(p_1, p_2,...., p_n)$.

(ii) $H(P)$ is continuous function of $(p_1, p_2,...., p_n)$ i.e. small change in $p_i's$, $i = 1,2,...,n$ should cause a small change in $H(P)$.

(iii) $H(P)$ should not change when $p_i's$, $i = 1,2,...,n$.

(iv) $H(P)$ is maximum when $p_1 = p_2 = ... = p_n = 1/n$.

(v) $H(1/n, 1/n, ..., 1/n)$ is monotonic increasing function of n.

(vi) $H(p_1, p_2, ..., p_n) = H(p_1 + p_2, p_3, ..., p_n) + (p_1 + p_2) H(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$.

Shannon did not state the first, third and fourth properties explicitly but these were implicit in what he stated. On the basis of these properties Shannon arrived at a measure

$$H(P) = -k \sum_{i=1}^{n} p_i \log p_i, \qquad p_i \geq 0, \forall i = 1,2,...,n, \tag{1.2.1}$$

where k is an arbitrary positive constant. This satisfies all the properties. Shannon also showed these properties characterized this measure and if all these properties are to be satisfied, and then this is the only function available.

When this measure was discovered by Shannon he did not want to call it "Information" since these words were already overworked. Therefore, he sought the advice of the great mathematician and physicist John Von Neumann, who advised that he should call it "ENTROPY" for two reasons. First, the function is already in use in thermodynamics under this name. Second and more importantly most people do not know what entropy really is and also if he uses the word entropy in an argument he will win every time Tribus [128]. The word entropy was introduced by Clausius in 1864 to explain the phenomenon of irreversibility associated with the study of real heat engines undertaken in 1824 by Carnot. He stated that second law of thermodynamics as the entropy law. Even Eddington stated: "The law that entropy always increases i.e., the second law of thermodynamics hold. I think supreme position among the laws of nature". The aim of Shannon was not to discuss thermodynamic entropy. He was only interested in information theoretic entropy. He just wanted to provide a quantitative measure for uncertainty, since uncertainty is associated with every probability distribution. Later Kapur (1972) has shown that thermodynamic entropy can be considered as the maximum information theoretic entropy of Shannon when average energy of physical system is prescribed.

1.2.2 Generalized Information Measures

Shannon's entropy was generalized in many ways according to researchers requirement.

1.2.2.1 Entropies of Order α and Order (α, β)

A systematic attempt to develop generalizations of Shannon's entropy was carried out by Rényi [95], who characterized entropy of order α given by

$$H_\alpha(P) = \frac{1}{1-\alpha}\log\sum_{i=1}^{n} p_i^\alpha, \quad \alpha \neq 1,\ \alpha > 0, \qquad (1.2.2)$$

where α is real parameter. We can easily verify that Eq. (1.2.2) reduces to Shannon entropy defined by Eq. (1.2.1) as $\alpha \to 1$.

Based on the same motivations of Rényi, later researchers (Aczél and Daróczy [2], Varma [133], Kapur [51], Rathie [93], etc.) generalized the entropy of order α by changing some of its postulates. The generalization studied by Aczél and Daróczy [2] known as entropy of order (α, β) is given by

$$H_{\alpha,\beta}(P) = \frac{1}{\alpha-\beta}\log\left(\frac{\sum_{i=1}^{n} p_i^\alpha}{\sum_{i=1}^{n} p_i^\beta}\right), \quad \alpha \neq \beta, \alpha > 0,\ \beta > 0, \qquad (1.2.3)$$

where α and β are real parameters.

1.2.2.2 Entropy of Degree α and Degree (α, β)

Havrda and Charvát [39] proposed the following entropy of degree α

$$H^\alpha(P) = (2^{1-\alpha}-1)^{-1}\left[\sum_{i=1}^{n} p_i^\alpha - 1\right], \quad \alpha \neq 1,\ \alpha > 0, \qquad (1.2.4)$$

for all $P = (p_1, p_2, ..., p_n) \in \Delta_n$. In this case, we can also verify that Eq. (1.2.4) reduces to Shannon entropy defined by Eq. (1.2.1) as $\alpha \to 1$.

Sharma and Taneja [107] studied a generalization of Eq. (1.2.4) involving two scalar parameters, known entropy of degree, (α, β) and is given by

$$H^{\alpha,\beta}(P) = (2^{1-\alpha} - 2^{1-\beta})^{-1} \sum_{i=1}^{n} (p_i^{\alpha} - p_i^{\beta}), \quad \alpha \neq \beta, \quad \alpha, \beta > 0, \tag{1.2.5}$$

for all $P = (p_1, p_2, ..., p_n) \in \Delta_n$, where α and β are real parameters.

In particular, when $\alpha = 1$ or $\beta = 1$, the measure given by Eq. (1.2.5) reduces to Eq. (1.2.4). In the limiting case, we have

$$\lim_{\alpha \to \beta} H^{\alpha,\beta}(P) = -2^{\alpha-1} \sum_{i=1}^{n} p_i^{\alpha} \log p_i, \quad \alpha > 0,$$

it reduces to Shannon entropy for $\alpha \to 1$.

1.2.2.3 Entropy of kind t

Arimoto [6] came up to a generalized entropy involving a real parameter, here we call it, entropy of kind t, given by

$$H_t(P) = (2^{t-1} - 1)\left[\left(\sum_{i=1}^{n} p_i^{1/t}\right)^{t} - 1\right], \; t \neq 1, \; t > 0, \tag{1.2.6}$$

for all $P = (p_1, p_2, ..., p_n) \in \Delta_n$. In this case also, we can easily verify that Eq. (1.2.6) reduces to Shannon entropy as $t \to 1$.

1.2.2.4 Entropies of Order 1 and Degree *β* and Order *α* and Degree *β*.

Sharma and Mittal [101] introduced and characterized two entropies called entropy of order 1and degree β and entropy of order α and degree β given by

$$H_1^{\beta} = (2^{1-\beta} - 1)^{-1}\left[\exp_2\left((\beta - 1)\sum_{i=1}^{n} p_i \log p_i\right) - 1\right], \; \beta \neq 1 \tag{1.2.7}$$

and

$$H_{\alpha}^{\beta} = (2^{1-\beta} - 1)^{-1}\left[\left(\sum_{i=1}^{n} p_i^{\alpha}\right)^{\frac{\beta-1}{\alpha-1}} - 1\right], \; \alpha \neq 1, \; \beta \neq 1, \; \alpha > 0 \tag{1.2.8}$$

respectively, for all $P = (p_1, p_2, ..., p_n) \in \Delta_n$, where α and β are real parameters.

Sharma and Mittal's main motivation was to generalize the three entropies, $H_\alpha(P)$, $H^\beta(P)$ and $H_t(P)$. With this aim, they arrived at $H_\alpha^\beta(P)$. The measure $H_\alpha^\beta(P)$ reduces to $H^\beta(P)$ and $H_t(P)$, when $\alpha = \beta$ and $\alpha^{-1} = t = 2 - \beta$, respectively. $H_\alpha^\beta(P)$ reduces to ${H_1}^\beta(P)$ and $H_\alpha(P)$, when $\alpha \to 1$ and $\beta \to 1$ respectively. Also, ${H_1}^\beta(P)$ reduces to Shannon's entropy, H(P), when $\beta \to 1$.

Thus, we see that the entropy of order α and degree β contain, either as a limiting or as a particular case, the Shannon's entropy, the entropy of order α, the entropy of degree β, the entropy of kind *t*, and the entropy of order *1* and degree β.

1.2.3 Relative Information and Inaccuracy

Kullback and Leibler's [68] measure of information associated with the probability distributions P and Q is given by

$$D(P\|Q) = \sum_{i=1}^{n} p_i \log(p_i / q_i). \tag{1.2.9}$$

The measure given by Eq. (1.2.9) has many names given by different authors such as, relative information, directed divergence, cross entropy, function of discrimination etc. Here we shall refer it relative information. It has found many applications in setting important theorems in Information theory and Statistics.

The Kerridge's [66] measure of information generally referred as inaccuracy associated with two probability distributions is given by

$$H(P\|Q) = \sum_{i=1}^{n} p_i \log q_i \,. \tag{1.2.10}$$

1.2.4 Divergence Measures

We see that the measure given by Eq. (1.2.9) is not symmetric in P and Q. Its symmetric version known as J-divergence (Jeffreys [48], Kullback and Leibler [68]) is given by

$$J(P\|Q) = D(P\|Q) + D(Q\|P) = \sum_{i=1}^{n} (p_i - q_i) \log\left(\frac{p_i}{q_i}\right). \qquad (1.2.11)$$

Sibson [109] for the first time introduced the idea of information radius for a measure arising due to concavity property of Shannon's entropy. This measure referred as Jensen difference divergence measure is given by

$$I(P \| Q) = H\left(\frac{P+Q}{2}\right) - \frac{H(P) + H(Q)}{2}$$

$$= \sum_{i=1}^{n} \left[\frac{p_i \log p_i + q_i \log q_i}{2} - \left(\frac{p_i + q_i}{2}\right) \log\left(\frac{p_i + q_i}{2}\right) \right]. \qquad (1.2.12)$$

By simple calculations, one can also write

$$I(P \| Q) = \frac{1}{2}\left[D\left(P \| \frac{P+Q}{2}\right) + D\left(Q \| \frac{P+Q}{2}\right) \right]. \qquad (1.2.13)$$

Taneja [122] studied an another kind of measure based on arithmetic and geometric mean inequality calling arithmetic and geometric mean divergence measure given by

$$T(P \| Q) = \frac{1}{2}\left[D\left(\frac{P+Q}{2} \| P\right) + D\left(\frac{P+Q}{2} \| Q\right) \right]$$

$$= \sum_{i=1}^{n} \left(\frac{p_i + q_i}{2}\right) \log\left(\frac{p_i + q_i}{2\sqrt{p_i q_i}}\right). \qquad (1.2.14)$$

Interestingly these three measures satisfy the following inequality:

$$I(P \| Q) + T(P \| Q) = 4J(P \| Q).$$

1.2.5 Weighted Entropies and their Generalization

In 1971, S. Guiasu [33] introduced the idea of Weighted Entropy by considering the utility importance of the events corresponding to its occurrence, viz.

$$H(P;U) = -\sum_{i=1}^{n} u_i p_i \log p_i, \quad p_i \geq 0, \sum_{i=1}^{n} p_i = 1, u_i > 0. \qquad (1.2.15)$$

Sharma-Mohan-Mitter [102] also considered utility of the event and studied independently Eq. (1.2.15) and called it 'Useful' measure of information and generalized Eq. (1.2.15) 'Useful' information/entropy of type β as follows:

$$H^{\beta}(P;U)=\frac{\sum_{i=1}^{n}u_i p_i(p_i^{\beta-1}-1)}{\sum_{i=1}^{n}u_i p_i},\quad \beta\neq 1,\ \beta>0. \tag{1.2.16}$$

For incomplete probability distributions, Picard [89] generalized these information measures by considering idea of preference and called non–additive information measures with preference of type β and of order α.

$$I_M^{\beta}(P;V)=\frac{1}{2^{1-\beta}-1}\left[2^{(1-\beta)\left(\sum_{i=1}^{n}v_i\log\frac{1}{p_i}\right)/\sum_{i=1}^{n}v_i}-1\right],\quad \beta\neq 1,\ \beta>0, \tag{1.2.17}$$

$$I_M^{\alpha,\beta}(P;V)=\frac{1}{2^{1-\beta}-1}\left[\left\{\frac{\sum_{i=1}^{n}p_i^{\alpha-1}v_i}{\sum_{i=1}^{n}v_i}\right\}^{\frac{\beta-1}{\alpha-1}}-1\right],\quad \alpha\neq 1,\beta\neq 1,\ \alpha\neq\beta,\ \alpha,\beta>0. \tag{1.2.18}$$

Sharma and Singh [104] studied generalized information measures with preference corresponding to Sharma and Taneja [107] generalized information measures

$$H_1^{\alpha}(P;U)=-2^{1-\alpha}\sum_{i=1}^{n}u_i p_i\log p_i,\ \alpha\neq 1,\ \alpha>0, \tag{1.2.19}$$

$$H^{\alpha,\beta}(P;U)=(2^{1-\alpha}-2^{1-\beta})^{-1}\sum_{i=1}^{n}u_i(p_i^{\alpha}-p_i^{\beta}),$$

for $\alpha\neq\beta,\ \alpha,\ \beta>0,\ \alpha\neq 1\neq\beta$ (1.2.20)

and

$$H_s^{\alpha,\beta}(P;U)=-\frac{2^{1-\alpha}}{\sin\beta}\sum_{i=1}^{n}u_i p_i^{\alpha}(\sin\beta\log p_i),$$

$$\alpha\neq\beta,\ \alpha,\ \beta>0,\ \alpha\neq 1\neq\beta. \tag{1.2.21}$$

Hooda-Tuteja [44] studied some generalization for incomplete probability distributions such as

$$H_{\alpha}{}^{\beta}(P;U) = (2^{1-\alpha} - 2^{1-\beta})^{-1} \frac{\sum_{i=1}^{n} u_i^{\alpha} p_i^{\alpha} (p_i^{\beta-\alpha} - 1)}{\sum_{i=1}^{n} p_i}, \alpha \neq \beta \quad (1.2.22)$$

1.2.6 Characterizations of Generalized Entropies

1.2.6.1 Entropy of Order α

The following characterization is due to Rényi [95]. Let us consider the following postulates defined for the function $H(P) \in \delta_n$, where

$$\delta_n = \left\{ P = (p_1, p_2, ..., p_n),\ p_i \geq 0,\ \sum_{i=1}^{n} p_i \leq 1 \right\}.$$

(i) $H(P)$ is a symmetric function of the elements of p.

(ii) If $\{p\}$ denotes the generalized probability distribution consisting of the single probability $\{p\}$ then $H(\{p\})$ is a continuous function of p in the interval $0 < p \leq 1$.

(iii) $H\left(\left\{\frac{1}{2}\right\}\right) = 1$.

(iv) For $P \in \delta_n$, $Q \in \delta_m$ and $P * Q \in \delta_{nm}$, we have

$$H(P * Q) = H(P) + H(Q).$$

Before stating the last postulate, we introduce some notations. Let $P = (p_1, p_2, ..., p_n) \in \delta_n$, and $Q = (q_1, q_2, ..., q_n) \in \delta_m$ be two generalized probability distributions such that $w(P) + w(Q) \leq 1$, we have

$$H(P \cup Q) = g^{-1}\left(\frac{w(P)g(H(P)) + w(Q)g(H(Q))}{w(P) + w(Q)} \right),$$

with $w(P) = \sum_{i=1}^{n} p_i \leq 1$, where g is strictly monotonic function.

Then

$$H(P) = H_\alpha(P) = \frac{1}{1-\alpha} \log \left(\frac{\sum_{i=1}^{n} p_i^\alpha}{\sum_{i=1}^{n} p_i} \right), \quad \alpha \neq 1,\ \alpha > 0 .$$

1.2.6.2 Entropy of Degree β

The following characterization of the measure given by Eq. (1.2.7) is due to Havrda and Charvát [39].

A function $H_n(p_1,...,p_n;\ \beta)$ will be said structural β-entropy if

(i) $H_n(p_1,...,p_n;\beta)$ is continuous in the region $p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1,\ \beta > 0$.

(ii) $H(1,\beta) = 0\,;\ H(\frac{1}{2},\frac{1}{2};\ \beta)\ =\ 1$.

(iii) $H_n(p_1,...,p_{i-1,}\ 0,\ p_{i+1},...,p_n;\beta)$

$= H_{n-1}(p_1,...,p_{i-1},\ p_{i+1},...,p_n;\beta),\ i = 1,2,...,n$.

(iv) $H_{n+1}(p_1,...,p_{i-1,}\ v_{i_1}, v_{i_2},\ p_{i+1},...,p_n;\beta)$

$= H_n(p_1,...,p_{i-1,}\ p_i,\ p_{i+1},...,p_n;\beta)\ + c p_i^\beta\ H_2(v_{i_1}/p_i, v_{i_2}/p_i\ ;\ \beta)$

$$v_{i_1} + v_{i_2} = p_i > 0,\ \ i = 1,2,...,n, \quad c > 0 .$$

Axioms (i)-(iv) determine the structural β-entropy given by Eq. (1.2.7).

Daróczy [27] presented an alternative way to characterize the entropy of degree β. An alternative way to characterize the entropy of degree β can also be seen in Sharma and Taneja [107].

1.2.6.3 Entropy of Degree (α, β)

Sharma and Taneja [107] presented an axiomatic characterization of degree (α, β). It is as follows:

Let $H_n^{\alpha,\beta} : \Delta_n \to R$ be a real valued function satisfying

$$H_n^{\alpha,\beta}(P) = \sum_{i=1}^{n} h(p_i) ,$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{m} h(p_i q_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_i^{\alpha} h(q_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} q_j^{\beta} h(p_i)\,, \qquad (1.2.23)$$

where $h:[0,1]\to R$ be a continuous function with $h\left(\frac{1}{2}\right)=\frac{1}{2}$. Then

$$H_n^{\alpha,\beta}(P) = (2^{1-\alpha} - 2^{1-\beta})^{-1} \sum_{i=1}^{n} (p_i^{\alpha} - p_i^{\beta}),\ \alpha \neq \beta,\ \alpha,\ \beta > 0\,. \qquad (1.2.24)$$

Sharma and Taneja [106] extended the functional Eq. (1.2.23) by the following generalized additivity

$$H(P * Q) = G(P)H(Q) + H(P)G(Q)\,,$$

for all $P \in \Delta_n,\ Q \in \Delta_m$ and $P * Q \in \Delta_{nm}$,

where $H(P) = \sum_{i=1}^{n} h(p_i)$ and $G(P) = \sum_{i=1}^{n} g(p_i)$, with h and g real valued continuous functions defined over [0,1] and $h\left(\frac{1}{2}\right)=\frac{1}{2}$. This lead us to the measure given by Eq. (1.2.5).

1.2.6.4 Entropy of Kind *t*

To obtain entropy of kind *t*, Arimoto [6] used a different approach. Let $f(u)$ be a real valued scalar function defined and nonnegative on (0, 1] with a continuous derivative on (0, 1] and $f(1)=0$. For all $P \in \Delta_n$, let us define

$$H_f(P) = H_f(p_1, p_2, ..., p_n) = \inf_{Q} \sum_{i=1}^{n} p_i f(q_i)\,, \qquad (1.2.25)$$

where the *inf.* is taken over all probability distributions such that

$$Q = (q_1, q_2, ..., q_n),\ \sum_{i=1}^{n} q_i,\ q_i > 0,\ \forall\ i = 1,2,...,n\,.$$

The function $H_f(P)$ given by Eq. (1.2.25) satisfy some interesting properties. These are summarized in the following properties:

(i) $H_f(p_1, p_2, ..., p_n)$ is a continuous and symmetric function with respect to its arguments $p_1, p_2, ..., p_n$.

(ii) $H_f(p_1, p_2, ..., p_n) = H_f(p_1, p_2, ..., p_n, 0)$.

(iii) $H_f(P)$ is a concave function with respect to P in Δ_n.

(iv) $0 \le H_f(p_1, p_2, ..., p_n) \le f(\frac{1}{n})$.

(v) If $f(u)$ is convex, then $H_f(p_1, p_2, ..., p_n) \le H_f(\frac{1}{n}, ..., \frac{1}{n}) \le f(\frac{1}{n})$.

(vi) In general, $H_f(p_1, p_2, ..., p_n) \le \sum_{i=1}^{n} p_i f(p_i)$.

(vii) If $f'(u) < 0$ on (0, 1), then

(a) $H_f(P) \ge H_f(H_o(P), 1 - H_o(P))$.

(b) $H_f(P) \ge f(\frac{1}{2}) H_o(P)$.

(viii) If $f(u)$ is convex with $f'(u) < 0$ on (0, 1], then

$$H_f(P) \ge f\left(1 - \frac{H_o(P)}{2}\right),$$

where $H_o(P) = 1 - \max\{p_1, p_2, ..., p_n\}$.

Let

$$f^t(u) = \begin{cases} (2^{t-1} - 1)^{-1}(u^{1-t} - 1), & t \ne 1,\ t \ge 0 \\ -\log_2 u, & t = 1 \end{cases}$$

then from Eq. (1.2.25), one has

$$H_t(P) = \begin{cases} 1 - \max\{p_1, p_2, ..., p_n\}, & t = 0 \\ (2^{t-1} - 1)^{-1}\left[\left(\sum_{i=1}^{n} p_i^{1/t}\right)^t - 1\right] & t \ne 1,\ t > 0 \\ -\sum_{i=1}^{n} p_i \log p_i, & t = 1 \end{cases}$$

The function $H_t(P)$ is the entropy of kind t. It arises as a particular case of Eq. (1.2.25). Boeeke and Lubbe [19] studied extensively the entropy of kind t by naming R-norm, considering $t = \frac{1}{R}$.

1.2.6.5 Entropy of Order α and Degree β

Sharma and Mittal [101] presented an axiomatic characterization of entropy of order α and degree β. It is based on the Rényi's approach, where the additivity property has been changed (generally referred as nonadditivity).

Let $H_n^{\beta}:\Delta_n \to R$ be a real valued continuous function satisfying

$$H(p_1, p_2,..., p_n) = g^{-1}\left(\frac{\sum_{i=1}^{n} p_i g(H(\{p_i\}))}{\sum_{i=1}^{n} p_i}\right),$$

where g is a strictly monotonic continuous function, and $H(\{P\})$, $0 < p \le 1$ is the self-information of an event of a probability distribution P satisfying

(i) $H(\{P\})$ is a continuous function of p in (0,1].

(ii) $H(\{Pq\}) = H(\{p\}) + H(\{q\}) + \lambda H(\{p\})H(\{q\}),\ \lambda \neq 0$.

(iii) $H(\{\frac{1}{2}\}) = 1$.

Then

$$H(p_1, p_2,...p_n;1,\beta) = (2^{1-\beta} - 1)^{-1}\left[\exp_2(\beta - 1)\frac{\sum_{i=1}^{n} p_i \log_2 p_i}{\sum_{i=1}^{n} p_i} - 1\right],\ \beta \neq 1, \beta > 0 \tag{1.2.26}$$

and

$$H(p_1, p_2,...p_n;\alpha,\beta) = (2^{1-\beta} - 1)^{-1}\left[\left(\frac{\sum_{i=1}^{n} p_i{}^{\alpha}}{\sum_{i=1}^{n} p_i}\right)^{\frac{\beta-1}{\alpha-1}} - 1\right],\ \alpha \neq, \beta \neq 1, \alpha > 0, \beta > 0. \tag{1.2.27}$$

Van der Pyl [132] restructured the above axiomatic system and considered as follows

Let $H_n:\Delta_n \to R$ be a real valued continuous function satisfying the following

(i) $H_n(P)$ is a symmetric function of its arguments.

(ii) $H_1(\{P\})$ is continuous in (0,1].

(iii) $H_1(\{\frac{1}{2}\}) = 1$.

(iv) There is a sequence $\{f_n\}$ such that

$$H_{nm}(P * Q) = H(P) + f_n(p_1, p_2, ..., p_n) H_m(Q),$$

for all $P \in \Delta_n$, $Q \in \Delta_m$ and $P * Q \in \Delta_{nm}$.

(v) There exists a strictly monotonic continuous function g such that

$$g(H_n(P)) = \left(\frac{\sum_{i=1}^{n} p_i g(H(\{p_i\}))}{\sum_{i=1}^{n} p_i} \right).$$

Then the above set of axioms leads to the measures given by Eq. (1.2.7) and Eq. (1.2.8).

1.2.7 Coding Theory

Coding Theory is one of the most important and direct applications of Information Theory. Coding theory is the study of the properties of codes and their fitness for a specific application. Codes are used for data compression, cryptography, error-correction and more recently also for network coding. Using a statistical description for data, Information Theory quantifies the number of bits needed to describe the data, which is the information entropy of the source.

There are essentially two aspects to coding theory:

1. Data compression or source coding.
2. Error correction or channel coding.

These two aspects may be studied in combination. Source encoding attempts to compress the data from a source in order to transmit it more efficiently. This practice is found every day on the Internet where the common Zip data compression is used to reduce the network load and make files smaller. The second, channel encoding, adds extra data bits to make the transmission of data more robust to disturbances present on the transmission channel.

A lot of coding theorems have been proved by many researchers based on the generalized information measures.

1.2.8 Inequalities

Inequalities in Information Theory have played a vital role in characterizing different information measures.

1.2.8.1 Shannon's Inequality

The very first inequality known as Shannon"s Inequality

i.e., $$-\sum p_i \log q_i \geq -\sum p_i \log p_i \,, \tag{1.2.28}$$

has played a vital role in coding theory. The inequality on right is the Shannon's entropy and on the left hand side, it is Kerridge [66] inaccuracy. It has been generalized in terms of functions such as

$$\sum p_i f(q_i) \leq \sum p_i f(p_i) \,, \tag{1.2.29}$$

of which the solution is

$$f(P) = a \log p + b \,. \tag{1.2.30}$$

1.2.8.2 Jensen's Inequality

$$\sum_{i=1}^{n} p_i q_i^w \begin{cases} \leq \left(\sum_{i=1}^{n} p_i q_i \right)^w, & 0 < w < 1 \\ \geq \left(\sum_{i=1}^{n} p_i q_i \right)^w, & w > 1 \ or \ \ w < 0 \end{cases} \tag{1.2.31}$$

with equality iff for some c, $q_i^w = c b_i^{w/w-1}$, $\forall i$, where q_i and b_i are non negative real numbers, for $w < 0,\ q_i > 0,\ b_i > 0,\ \forall i$.

1.2.8.3 Holder's Inequality

$$\left(\sum_{i=1}^{n} p_i^{w}\right)^{1/w}\left(\sum_{i=1}^{n} p_i^{\frac{w}{w-1}}\right)^{\frac{w-1}{w}}\begin{cases}\leq \sum_{i=1}^{n} p_i q_i, & w<1,\ \ w\neq 0\\ \geq \sum_{i=1}^{n} p_i q_i, & w>1\end{cases} \tag{1.2.32}$$

with equality iff for some c, $p_i^{w} = cq_i^{w/w-1}$, $\forall i$, where p_i and q_i are non negative real numbers, for $w<0,\ q_i>0,\ p_i>0,\ \forall i$.

1.2.8.4 General Inequality

$$\left(\sum_{i=1}^{n} p_i\right)^{w}\begin{cases}\leq \sum_{i=1}^{n} p_i{}^{w}, & w<1,\ w\neq 0\\ \geq \sum_{i=1}^{n} p_i{}^{w} & w>1\end{cases} \tag{1.2.33}$$

1.3 Survey of Literature

In 1924 the first studies in Information Theory were undertaken by Nyquist [87], and in 1928 by Hartley [38] who recognized the logarithmic nature of the measure of information. In 1948, Shannon [98] published a remarkable paper on the properties of information sources and of the communication channels used to transmit the outputs of these sources. Around the same time Wiener [134] also considered the communication situation and came up, independently, with results similar to those of Shannon.

Both Shannon and Wiener considered the communication situation as one in which a signal, chosen from a specified class, is to be transmitted through a channel. The output of the channel is described statistically by each permissible input. The basic problem of communication is to reconstruct as closely as possible the input signal after observing the received signal at the output. However, the approach used by Shannon differs from that of Wiener, in the

nature of the transmitted signal and in the type of decision made at the receiver. In the Shannon model messages are first encoded and then transmitted, whereas in the Wiener model the signal is communicated directly through the channel without being encoded.

A key feature of Shannon Information Theory is the term "information" that can often be given a mathematical meaning as a numerically measurable quantity, on the basis of a probabilistic model, in such a way that the solutions of many important problems of information storage and the transmission can be formulated in terms of this measure of the amount of information. This important measure has a very concrete operational interpretation: it is roughly equals the minimum number of binary digits needed, on the average, to encode the message in question. The coding theorems of Information Theory provide such overwhelming evidence for the adequateness of the Shannon information measure that to look for essentially different measures of information might appear to make no sense at all. Moreover, it has been shown by several authors, starting with Shannon [98], that the measure of amount of information is uniquely determined by some rather natural postulates. Still, all the evidence for which Shannon information measure is possible, is valid only within restricted scope of coding problems considered by Shannon. As pointed out by Rényi [95] in his fundamental paper on generalized information measures, in some of the problems other quantities may serve just as well, or even better, as measures of information. This should be supported either by their operational significance or by a set of natural postulates characterizing them or preferably by both. Thus the idea of generalized entropies arises in the literature. It started with Rényi [95] who characterized a scalar parametric entropy as entropy of order α, which includes Shannon entropy as a limiting case.

On the other side, Kullback and Leiber [68] studied a measure of information from statistical aspects of view, involving two probability distributions associated with the same experiment, calling discrimination function, later different authors named as cross entropy, relative information, etc. At the same time Kullback and Leiber also studied a divergence measure, calling J-divergence, the measure already studied by Jeffreys [48]. Kerrige [66] studied a different kind of measure,

calling inaccuracy measure, involving again two probability distributions. Sibson [109] studied another divergence measure involving two probability distributions, using mainly the concavity property of Shannon's entropy, called information radius. Later Burbea and Rao [22] studied extensively the information radius and its parametric generalization, called this measure as Jensen difference divergence measure. Thus, the Shannon's entropy, the Kullback-Leibler's relative information, the Kerridge's inaccuracy, the Jeffreys invariant (or J-divergence) and Sibson's information radius are the five classical measures of information associated with one and two probability distributions. These five classical measures have found deep applications in the areas of Information Theory and Statistics. During the past years various measures have been introduced in the literature generalizing these measures and these generalizations include one and two scalar parameters. Taneja [122] studied a new measure of divergence and its two parametric generalizations involving two probability distributions based on arithmetic and geometric mean inequality.

Agarwal [4], Sharma, Mohan and Mitter [102], Hooda and Tuteja [40], Sharma and Singh [104, 105, 110, 111], Gurdial and Pesson [36], Singh and Bhardwaj [112, 113], Singh and Taneja [118], Omprakash [91, 92], Kumar and Choudhary [69, 70, 71, 72, 73, 74, 75, 76, 77, 78] have done a lot of work in this direction.